An improved method for predicting the accuracy of genetic stock identification

Eric C. Anderson, Robin S. Waples, and Steven T. Kalinowski

Abstract: Estimating the accuracy of genetic stock identification (GSI) that can be expected given a previously collected baseline requires simulation. The conventional method involves repeatedly simulating mixtures by resampling from the baseline, simulating new baselines by resampling from the baseline, and analyzing the simulated mixtures with the simulated baselines. We show that this overestimates the predicted accuracy of GSI. The bias is profound for closely related populations and increases as more genetic data (loci and (or) alleles) are added to the analysis. We develop a new method based on leave-one-out cross validation and show that it yields essentially unbiased estimates of GSI accuracy. Applying both our method and the conventional method to a coastwide baseline of 166 Chinook salmon (*Oncorhynchus tshawytscha*) populations shows that the conventional method provides severely biased predictions of accuracy for some individual populations. The bias for reporting units (aggregations of closely related populations) is moderate, but still present.

Résumé : L'estimation de la précision de l'identification du stock génétique (« GSI ») qu'on peut espérer, étant donné une banque de données de base récoltée antérieurement, nécessite des simulations. La méthode courante comprend des simulations répétées de mélanges par ré-échantillonnage de la banque de données de base, des simulations de nouvelles banques de données de base en ré-échantillonnant la banque de données et l'analyse des mélanges ainsi simulés à l'aide des banques de données de base simulées. Nous montrons que cette méthode surestime la précision prédite de GSI. L'erreur est importante dans les populations fortement apparentées et elle augmente à mesure que de nouvelles données génétiques (locus et (ou) allèles) sont ajoutées à l'analyse. Nous mettons au point une nouvelle méthode basée sur une validation croisée de type « leave-one-out » (avec retrait d'un élément) et nous montrons qu'elle produit essentiellement des estimations non erronées de la précision de GSI. L'application de notre méthode et de la méthode courante à une banque de données de base provenant de 166 populations de saumons chinook (*Oncorhynchus tshawytscha*) réparties sur toute la côte montre que la méthode courante fournit des prédictions de la précision qui sont grandement faussées pour certaines populations individuelles. L'erreur dans le cas des unités d'évaluation (des rassemblements de populations fortement apparentées) est peu importante, mais réelle.

[Traduit par la Rédaction]

Introduction

Genetic data have been used to estimate the stock composition of mixed-stock fisheries for over two and a half decades (e.g., Grant et al. 1980; Milner et al. 1981). The basic methodology for this is straightforward: fish from the mixed fishery are genotyped, as are fish in "baseline" samples, which are taken separately from the populations that might contribute to the mixture. Mixture proportions in the fishery are then estimated using conditional maximum likelihood (Milner et al. 1981; Fournier et al. 1984; Millar 1987), unconditional maximum likelihood (Smouse et al. 1990), or Bayesian (Pella and Masuda 2001, 2006) methods that relate the genotypes in the mixture to the expected genotype frequencies in the baseline populations. Most applications of this sort of genetic stock identification (GSI) have involved Pacific salmon, but the basic methodology has now been applied to a wide range of other species as well (Waldman et al. 1997; McParland et al. 1999; Koljonen et al. 2005).

Initially, allozyme polymorphisms used for GSI provided sufficient resolution to address many Pacific salmon management concerns; however, other types of genetic data are now becoming increasingly abundant and inexpensive. Highly polymorphic microsatellite loci have been shown to provide considerable power for GSI (Kalinowski 2004; Winans et al. 2004). Declining costs of genotyping and the recent completion of a cross-laboratory, standardized, microsatellite DNA baseline for over 150 populations throughout the range of Chinook salmon (Seeb et al. 2007) have led to microsatellites replacing allozymes for many GSI applications. In addition, numerous single nucleotide polymorphisms (SNPs) are being discovered in Pacific salmon, and these may soon become useful, low-cost, high-throughput genetic markers for GSI (Smith et al. 2005*a*). At the same

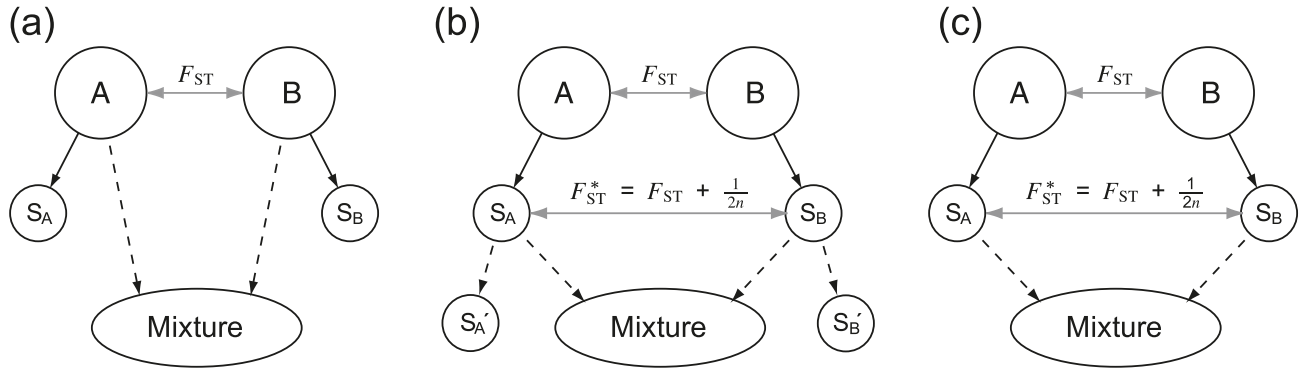
Received 16 May 2007. Accepted 31 October 2007. Published on the NRC Research Press Web site at cjfas.nrc.ca on 25 June 2008. J20006

E.C. Anderson.[1] Fisheries Ecology Division, Southwest Fisheries Science Center, 110 Shaffer Road, Santa Cruz, CA 95060, USA.
R.S. Waples. Northwest Fisheries Science Center, 2725 Montlake Boulevard East, Seattle, WA 98112, USA.
S.T. Kalinowski. Department of Ecology, 310 Lewis Hall, Montana State University, Bozeman, MT 59717, USA.

[1]Corresponding author (e-mail: eric.anderson@noaa.gov).

 doi:10.1139/F08-049 © 2008 NRC Canada

Fig. 1. Diagrams representing different ways of simulating mixtures for the assessment of genetic stock identification (GSI) accuracy with two populations (A and B) in the baseline. Both baseline samples, S_A and S_B, are assumed to be of size n individuals. Broken arrows represent the simulation of genotypes. (*a*) Individuals in the mixture are drawn directly from parametric population allele frequencies, as are the baseline samples used to analyze the mixture. The magnitude of difference in parametric allele frequencies between A and B is indicated by F_{ST}. Pursuing this method is typically impossible because the allele frequencies in A and B are never known without error. (*b*) The conventional method in use today: parametric bootstrap with baseline resampling (PB-R). The baseline sample provides estimates of allele frequencies, and these estimates are used to generate simulated individuals for the mixture and for a new "resampled" baseline. On average, the magnitude of allele frequency difference between the populations, estimated from the baseline (F^*_{ST}), will be larger than between the true population allele frequencies; $F^*_{ST} = F_{ST} + \frac{1}{2n}$. As a result, individuals in the mixture will be genetically more divergent than is really the case, leading to overly optimistic estimates of power. (*c*) As in *b*, but without baseline resampling (PB-NR).



time, there is a growing need for these technologies in fisheries management: attempts are being made not only to apply GSI to mixtures of evermore genetically similar stocks, but also to identify the origin of individual fish in mixed fisheries rather than just the aggregate proportion of fish from different populations. These management tasks require highly informative genetic data.

This confluence of management needs and data availability makes it imperative that fisheries geneticists have at their disposal a reliable method for assessing the expected accuracy of GSI with a given baseline data set. Developing such a method is the main purpose of this paper. GSI is most accurate when there is a high degree of genetic differentiation between populations, when the baseline samples for each population are large, and when large numbers of loci are genotyped (Wood et al. 1987; Kalinowski 2004), but there is no simple formula available to calculate how accurate GSI is expected to be; therefore computer simulation must be used. However, as we explain below, the simulation methods currently in use to do this are flawed in a way that leads them to consistently overestimate the expected accuracy of GSI.

If parametric allele frequencies in source populations could be known without error, then estimating the accuracy of GSI via computer simulation would be straightforward (Fig. 1*a*). Baseline and mixture samples could be simulated from the allele frequencies in the source populations, and these data could be used to estimate mixture proportions. The mean squared error (or some other summary statistic) for a large number of these estimates would provide an estimate of how accurate GSI is likely to be. In reality, parametric allele frequencies of source populations are never known exactly, so an alternative simulation procedure must be used. The standard approach in such situations is to assume that the parametric allele frequencies in the source populations are equal to the observed allele frequencies in the baseline samples taken from the source populations (Fig. 1*b*). Baseline and mixture genotypes are then simulated from these allele frequencies in the same manner as they are when parametric allele frequencies are known (Fournier et al. 1984; Beacham et al. 2006). The only difference is that the allele frequencies in the baseline samples are used as if they were the parametric allele frequencies.

This approach for estimating the accuracy of GSI has been in use for a long time in a variety of computer packages (including some of our own) and produces reasonable estimates of the accuracy of GSI in many circumstances; however, as we show below, GSI is generally less accurate than these conventional simulations predict it will be, and the discrepancy is most severe for the most challenging applications of GSI. The crux of the problem is that conventional simulation does not properly account for sampling error in baseline allele frequencies, and this error induces a spurious correlation between the baselines and the individuals simulated by resampling from the baseline. Sampling error makes the allele frequency divergence observed between samples larger, on average, than the true differences between the populations. Using these observed sample allele frequencies to naively drive simulations of GSI will lead one to exaggerate the accuracy of GSI. The magnitude of this problem can be understood heuristically in terms of F_{ST}, a standardized measure of genetic differentiation between populations that is familiar to most population geneticists. If sampling error is not accounted for, F_{ST} for two samples from a pair of populations will be inflated by a term of approximate magnitude $1/(2n)$ (Wright 1978; Chakraborty and Leimar 1987), where n is the number of individuals sampled. The estimated value of F_{ST}, F^*_{ST}, between two source populations will then be equal to the true value plus $1/(2n)$. Because the accuracy of GSI increases with F_{ST}, an inflated F^*_{ST} will lead to overly optimistic assessments of how accurate GSI is likely to be.

A few numerical examples illustrate the potential importance of this type of bias. Assume that the parametric F_{ST}

© 2008 NRC Canada

between source populations A and B is 0.02 and that the baseline samples, S_A and S_B, for these populations each include $n = 50$ individuals. The expected magnitude of inflation of F^*_{ST} due to sampling error is then $1/(2n) = 0.01$. In this case, the degree of genetic differentiation that characterizes individuals simulated for the mixture is $F^*_{ST} = 0.02$ (parametric) + 0.01 (baseline sampling) = 0.03, which is inflated by 50% compared with the true F_{ST} for the populations. In the extreme, if A and B are really parts of the same panmictic population, there is no basis whatsoever for resolving a "mixture" of A and B individuals. However, a mixture analysis conducted as in Fig. 1*b* would suggest that some ability to resolve this mixture does exist. Furthermore, as more and more loci are added to the analysis, the apparent power to resolve this spurious mixture will increase. If, on the other hand, allele frequency differences between populations actually are much larger (e.g., $F_{ST} = 0.1$), the mixture analysis will have high intrinsic power, and the relatively small inflation of F^*_{ST} due to baseline sampling will have comparatively little effect on estimated power.

In this paper, we propose a new method based on a leave-one-out procedure for assessing GSI accuracy that is easy to implement and that reduces the bias to a negligible amount in most cases. Leave-one-out cross validation is not new to genetic classification problems (Spielman and Smouse 1976; Piry et al. 2004). However, the utility of the leave-one-out procedure has not been fully appreciated in the context of GSI, perhaps because its application to GSI is not as straightforward or obvious as its application to genetic assignment tests. We illustrate the utility of the new method and evaluate the magnitude of the bias in the conventional method using simulated data. Finally, we assess GSI accuracy in a large microsatellite baseline using both the conventional biased method and our new method to demonstrate the practical implications of the bias.

Methods

Once a baseline sample, which includes, say, M populations, has been collected, it is natural to inquire how accurately that baseline can be used to estimate the mixed-stock fishery proportions, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$, using a fishery sample of size N fish that might be collected in the future. For this exercise, it is typical to assume some value $\boldsymbol{\pi}^*$ of the mixing proportions and perform simulations to see how closely the maximum likelihood estimates, $\widehat{\boldsymbol{\pi}}$, correspond to $\boldsymbol{\pi}^*$. The conventional method to do this, as implemented in SPAM (Alaska Department of Fish and Game 2000) and GMA (Kalinowski 2003), involves the following steps for each simulated data set. (1) The number of fish in the mixture from each of the M populations is drawn from a multinomial distribution of N trials with cell probabilities $\boldsymbol{\pi}^*$. (2) A genotype $\boldsymbol{G}$ for each fish in the mixture from population i is simulated by sampling, with replacement, from population i's allele frequencies, $\widehat{\boldsymbol{\theta}}_i$, as estimated from the observed baseline. (3) A new baseline sample of the same size as the observed baseline is created by sampling alleles, with replacement, from the observed baseline. A genotype for each fish in the simulated baseline from population i is simulated by sampling with replacement from $\widehat{\boldsymbol{\theta}}_i$. (4) The simulated baseline is applied, using the standard conditional-likelihood GSI framework (Fournier et al. 1984; Millar 1987), to find the conditional MLE, $\widehat{\boldsymbol{\pi}}$, from the simulated mixture.

The conditional likelihood function commonly in use today is an integrated likelihood that assumes a unit-information Dirichlet prior on allele frequencies (Rannala and Mountain 1997). In such a case, genotypes can be simulated from their posterior predictive distribution (a compound Dirichlet–multinomial distribution) rather than by simply sampling alleles with replacement from the baselines. There seems to be little practical difference between these two approaches.

Although the unconditional likelihood framework for GSI (Smouse et al. 1990) and Bayesian methods (Pella and Masuda 2001, 2006) may be preferable to a conditional likelihood analysis, the conditional likelihood framework is usually used for analyzing the simulated data sets because it is customary to analyze many simulated data sets, and maximizing the conditional likelihood requires far less computer time than maximizing the unconditional likelihood or computing the Bayesian posterior distribution.

We call the above method the parametric bootstrap method with baseline resampling (PB-R). A second method, which is available as an option in SPAM, is the parametric bootstrap with no baseline resampling (PB-NR). PB-NR is identical to PB-R except that step 3 is omitted and the simulated mixture is analyzed with the original baseline sample. Our simulations (see below) show that both PB-R and PB-NR provide upwardly biased predictions of GSI accuracy.

We introduce a slight modification to the steps above to provide an essentially unbiased method for predicting GSI accuracy. It proceeds as follows: step 1 is unmodified; step 2 is slightly modified, the two gene copies at a locus in a single individual are drawn without replacement (but they are both replaced before the next individual is simulated); step 3 is omitted (it is not necessary to simulate a new baseline sample); and in step 4, a leave-one-out procedure is used when computing $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i)$, the probability of the simulated fish's genotype $\boldsymbol{G}$ given the baseline allele frequencies. This leave-one-out procedure takes the following form: if i is not the population from which the individual was simulated, then $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i)$ is computed as before. If, however, i is the population from which the individual was simulated, then $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i^{(-)})$ is used instead of $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i)$. $\widehat{\boldsymbol{\theta}}_i^{(-)}$ is just the MLE of $\boldsymbol{\theta}_i$ computed after subtracting the gene copies carried in $\boldsymbol{G}$ from the genes found in the baseline from population i.

We refer to this new method as the cross-validation method over gene copies (CV-GC). We explore two other methods as well: cross validation over single locus genotypes (CV-SL) and over multilocus genotypes (CV-ML). CV-SL is identical to CV-GC except that in step 2, each single-locus genotype is simulated by randomly sampling, with equal probability, from among the single-locus genotypes carried by members of the appropriate population in the baseline. CV-ML is identical to CV-GC except that in step 2 each individual's full, multilocus genotype is simulated by randomly sampling, with equal probability, from among the multilocus genotypes carried by members of the appropriate population in the baseline.

Some notation will be useful for the following section: we

© 2008 NRC Canada

use $\widehat{\boldsymbol{T}}_j$ to denote the scaled likelihood vector of the jth individual in a mixture. If the jth individual has genotype $\boldsymbol{G}$, this vector is obtained by starting with a vector the elements of which are all $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i)$ and scaling them to sum to one. It shall be understood that when using any of the leave-one-out, cross-validation methods, one of the elements of $\widehat{\boldsymbol{T}}_j$ will have been computed using $P(\boldsymbol{G}|\widehat{\boldsymbol{\theta}}_i^{(-)})$. It can be shown by the factorization theorem (Casella and Berger 1990, p. 250) that the set of all $\widehat{\boldsymbol{T}}_j$s is the sufficient statistic for $\boldsymbol{\pi}$, and a mathematical justification for our cross-validation methods may be described in terms of the distribution of simulated values of this sufficient statistic; however, for brevity, we omit those details. The simulation results provide enough evidence of the utility of our new methods.

Simulation experiments

We perform three separate experiments using purely simulated data. In the first two, the structured coalescent model (Hudson 1990) implemented in *makesamples* (Hudson 2002) is used to simulate genotypes of individuals from different populations in both the baseline and the mixture samples. Each locus was simulated using a separate, independent realization of the coalescent process on a nonrecombining DNA segment that we will refer to as a "chromosome". This creates loci that are independently segregating. An island model of migration was used to effect a given degree of divergence between the populations. Under this model, populations that exchange many migrants each year are not very highly diverged, and populations that exchange few migrants each year will be more highly diverged. In *makesamples*, the migration rates are parameterized by a migration matrix $[4N_em_{i,j}]$, where $m_{i,j}$ is the fraction of population i that is composed of migrants from population j each generation and N_e is the effective size of each population. N_e was assumed to be equal among all populations and constant in time.

For example, to simulate 10 microsatellite loci in 200 individuals (400 chromosomes) from each of three populations in a symmetric island model with $4N_em_{i,j} = 5$, we would first use the *makesamples* command line "ms 1200 10 –t 8.0 –I 3 400 400 400 10". The above produces data using the infinite sites model of mutation with rate controlled by the parameter $4N_e\mu = 8.0$, where μ is the rate of neutral mutation each generation. Each mutation in the infinite-sites model corresponds to a mutation at a single, unique nucleotide on the chromosome. For each chromosome in the simulation, *makesamples* returns an ordered string of 1s and 0s, which denote whether the chromosome did or did not, respectively, inherit the mutated form of the nucleotide. We convert those ordered arrays of infinite-sites mutations to microsatellite alleles following a stepwise mutation model with occasional multistep mutations in the following manner: the ancestral microsatellite allele length is set to a large value and each mutant nucleotide in the locus is randomly assigned to be an allele length increase or decrease with equal probability. The magnitude of the length change associated with each mutant nucleotide with probability of 0.85 was chosen to be 1 and with probability of 0.15 was $2 + W$, where W is drawn from a Poisson distribution with a mean of 3. Our mutation parameters yielded simulated loci with an average of 15 alleles. The simulation of microsatellite alleles from infinite sites mutations was done using the program *ms2geno* available from E. Anderson upon request.

For each simulation replicate, the genotypes of a large number of individuals from each population were simulated. Some of those individuals were randomly assigned to be members of the baseline samples, and other, separate individuals were chosen to be members of the mixture samples. We simulated different numbers of individuals under varying parameter values in the two coalescent simulation experiments. Specifics are described below. The number of individuals in the baseline sample from population p is denoted by n_p, the number of individuals from population p in the mixture is denoted by N_p, and the size of the whole mixture sample is denoted by N.

All the simulation output was analyzed using the program *gsi_sim* written by Eric Anderson. The programs GMA, SPAM, and cBayes (Neaves et al. 2005) are not amenable to large-scale simulation studies because they are available only for the Microsoft Windows platform and cannot be easily scripted in a Unix-like environment. We verified the results of our large simulations on small test cases analyzed by GMA and SPAM. All of the analyses, and the parametric bootstrapping of mixtures, were done using the likelihood model proposed by Rannala and Mountain (1997) with the unit-information prior. Nearly identical results were obtained using the unifom prior (result not shown).

Our first experiment investigates the simplest population scenario possible: a mixture formed from two populations, indexed 1 and 2, simulated with a migration rate $4N_em_{1,2} = 4N_em_{2,1} \in \{\text{PANMIX}, 1000, 250, 100, 25, 10\}$ between them. We use "PANMIX" to denote the case in which the populations are entirely panmictic. It is sometimes easier to think of these migration rates in terms of the expected F_{ST} values that they will produce between the populations. Under an infinite alleles model, the expected F_{ST} values would be approximately 0, 0.001, 0.004, 0.01, 0.04, and 0.09, respectively. Mutation rate was $4N_e\mu = 4.5$ (for the PANMIX case, $4N_e\mu$ was set to 8.0 to ensure roughly the same number of alleles in all cases). Baseline sample sizes from each of the two populations were $n_1 = n_2 = 144$, and each individual had genotypes at $L = 13$ loci. These sample sizes and L are identical to those throughout the baseline gathered by the Pacific Salmon Commission (PSC) Genetic Analysis of Pacific Salmon (GAPS) Consortium (Seeb et al. 2007). We compared the true distributions of $\widehat{\boldsymbol{T}}_j$ and $\widehat{\boldsymbol{\pi}}$ with the distributions obtained by simulating new individuals using resampling from the baseline allele frequencies. The true distributions, which we refer to as "true", were obtained by computing $\widehat{\boldsymbol{T}}_j$ and estimating $\widehat{\boldsymbol{\pi}}$ from mixtures containing individuals that were distinct from the individuals in the baselines, but which were simulated on the same coalescent trees. The distributions of $\widehat{\boldsymbol{T}}_j$ and $\widehat{\boldsymbol{\pi}}$ for mixtures formed by resampling from the baseline allele frequencies were obtained using the PB-NR, PB-R, CV-ML, CV-SL, and CV-GC methods described above. Mixtures of $N = 400$ were simulated. For assessing the distributions of $\widehat{\boldsymbol{T}}_j$, $N_1 = 400$ and $N_2 = 0$ and we investigated the distribution of the mean over all fish j of $\widehat{T}_{j,1}$. The distribution of $\widehat{\boldsymbol{\pi}}$ was assessed in the simulation with $N_1 = 400$ and another with $N_1 = 250$. Note

© 2008 NRC Canada

Table 1. Comparison of the CV-GC and PB-R methods using the GAPS microsatellite baseline (version 2.1 of the baseline, unpublished data, obtained in 2006 from M. Banks, Coastal Oregon Marine Experiment Station, Hatfield Marine Science Center, Department of Fisheries and Wildlife, Oregon State University, Newport, OR 97365) and mixing proportions from the 2006 Monterey Bay recreational fishery.

			Mean π_i		$\sqrt{\text{MSE of } \pi_i}$		Fraction of individual assignments correct	
Region or reporting unit	Population	True π	CV-GC	PB-R	CV-GC	PB-R	CV-GC	PB-R
Central Valley Fall	Feather Hatchery Fall	0.3298	0.3374	0.3441	0.0337	0.0377	0.504	0.746
Central Valley Spring	Feather Hatchery Spring	0.2731	0.2704	0.2742	0.0292	0.0303	0.596	0.773
Central Valley Fall	Battle Creek	0.2069	0.2062	0.2107	0.0265	0.0272	0.513	0.739
Central Valley Fall	Stanislaus River	0.0570	0.0526	0.0465	0.0168	0.0176	0.151	0.498
Central Valley Fall	Butte Creek Fall	0.0489	0.0467	0.0377	0.0156	0.0169	0.145	0.462
Central Valley Winter	Sacramento Hatchery	0.0321	0.0317	0.0320	0.0086	0.0088	0.999	1.000
Central Valley Spring	Butte Creek Spring	0.0150	0.0147	0.0148	0.0063	0.0062	0.905	0.947
Central Valley Spring	Mill Creek Spring	0.0083	0.0088	0.0086	0.0053	0.0049	0.552	0.748
Rogue River	Applegate Creek	0.0077	0.0063	0.0072	0.0038	0.0040	0.732	0.854
California Coast	Eel River	0.0039	0.0036	0.0037	0.0024	0.0024	0.896	0.952
California Coast	Russian River	0.0030	0.0032	0.0031	0.0024	0.0024	0.909	0.945
Klamath River	Trinity Hatchery Spring	0.0028	0.0026	0.0027	0.0023	0.0023	0.646	0.788
Northern California – southern Oregon Coast	Chetco River	0.0026	0.0025	0.0026	0.0020	0.0021	0.900	0.942
Klamath River	Klamath River Fall	0.0024	0.0023	0.0024	0.0021	0.0021	0.733	0.845
Klamath River	Trinity Hatchery Fall	0.0020	0.0023	0.0022	0.0024	0.0022	0.636	0.766
Central Valley Spring	Deer Creek Spring	0.0018	0.0019	0.0017	0.0023	0.0018	0.303	0.610
Upper Columbia River	Hanford Reach	0.0014	0.0012	0.0016	0.0019	0.0021	0.453	0.667
Rogue River	Cole Rivers Hatchery	0.0005	0.0012	0.0008	0.0020	0.0015	0.587	0.750
Mid-Oregon Coast	Umpqua Hatchery	0.0005	0.0009	0.0007	0.0017	0.0015	0.510	0.723
Mid-Oregon Coast	Coquille River	0.0005	0.0003	0.0004	0.0007	0.0009	0.557	0.736

Note: True π refers to the value of π used to generate the simulated mixtures. CV-GC, cross-validation method over gene copies; PB-R, parametric bootstrap with baseline resampling. Summary statistics for $\widehat{\pi}_i$ and the fractions of correct individual assignments were calculated from 2500 simulated mixed-fishery samples.

that in this case, the proportion of fish from the two populations is fixed in the sample itself. Thus the variance of $\widehat{\boldsymbol{\pi}}$ does not include the variance due to finite sampling of fish from the mixture. For each value of N_1, 1000 simulated baselines and mixtures were used. For analyses involving resampling, a single mixture was simulated and analyzed per baseline. Thus, each distribution is estimated using 1000 realizations of the variable.

In the second experiment, we simulated five populations in a symmetric island model with migration rates $4N_e m_{i,j} \in \{\text{PANMIX}, 62.5, 25, 6.25, 2.5\}$ for all $i \neq j$. In a single coalescent simulation for each migration rate, we generated enough individuals to produce 25 replicate baselines containing 144 individuals from each of the five populations and 250 mixture samples of size $N = 400$. The true mixing proportions $\boldsymbol{\pi}$ for populations 1–5 were 0.5, 0.25, 0.125, 0.0625, and 0.0625, respectively. We analyzed each mixture sample using the 25 replicate baselines. Additionally, using both the CV-GC and the PB-R methods, for each of the 25 baselines, we created and analyzed 250 mixtures by resampling genes from the baselines. For each of the 25 baselines, the average value of $\widehat{\boldsymbol{\pi}}$ over the 250 mixtures was recorded. The results are presented in terms of the mean and standard deviation of this mean $\widehat{\boldsymbol{\pi}}$ over the 25 baselines considered across the five different populations and the five different migration rates.

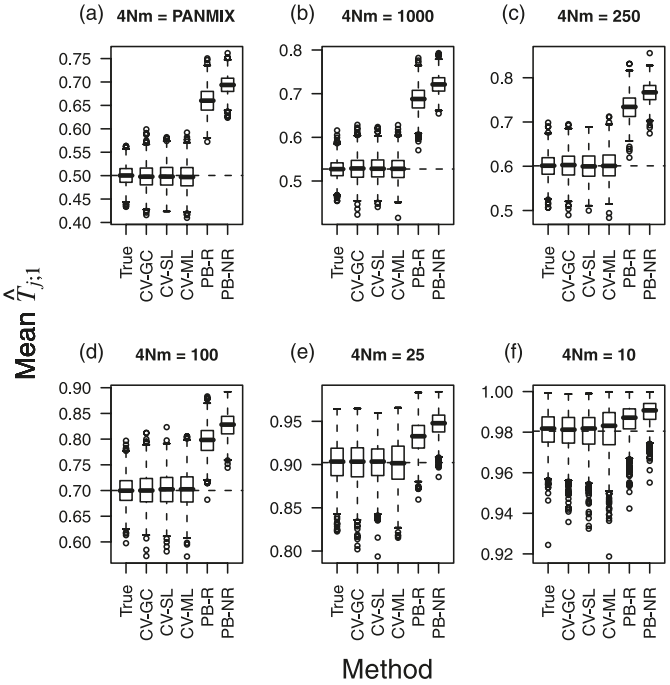
The third and final simulation was designed to determine whether the bias of the PB-R method is a complication of using loci with large numbers of alleles, some of which are at very low frequency, or whether the bias is evident even when the loci used have few alleles, all of them at intermediate frequency. Kalinowski (2004) found that the total number of alleles (across all loci) was a key determinant of the accuracy of GSI. Therefore, we fixed K_{tot}, the total number of informative alleles in the simulation, at 256, and then varied L, the number of loci, in powers of two from 1 to 256. We assumed two populations with identical, uniform, allele frequencies. Thus, for a simulation with L loci, each locus had $k = 1 + K_{\text{tot}}/L$ alleles, each at a frequency of $1/k$. Five-hundred different baselines were simulated by sampling 144 individuals for each population from these uniform allele frequencies. For each baseline, a single mixture of $N = 400$ fish, all of them from population 1, was simulated and analyzed using the CV-ML, CV-SL, CV-GC, PB-R, and PB-NR methods. For each mixture, we recorded the average value, over individuals in the mixtures, of $\widehat{T}_{j,1}$. We present the results in terms of the average value of $\widehat{T}_{j,1}$ over the 500 replicate baselines. Because the two populations have identical allele frequencies, the expected value of $\widehat{T}_{j,1}$ is 0.5, and any systematic departure from that is a sign of bias.

Analysis of empirical data

We use the coastwide Chinook microsatellite baseline developed by the PSC-funded GAPS consortium (version 2.1 of the baseline, unpublished data, obtained in 2006 from M. Banks, Coastal Oregon Marine Experiment Station, Hatfield Marine Science Center, Department of Fisheries and Wildlife, Oregon State University, Newport, OR 97365) as the

© 2008 NRC Canada

Fig. 2. Box plots of $\widehat{\boldsymbol{T}}_j$ simulated in five different ways. Bold lines are the medians, boxes span the interquartile range, whiskers extend to 1.5× the interquartile range beyond each box edge, and open circles denote data points falling beyond the whiskers. For each replicate, 400 fish from population 1 were simulated and the mean value of $\hat{T}_{j,1}$ over those 400 fish was computed. Each box plot summarizes the distribution of the mean of the $\hat{T}_{j,1}$ values over 1000 replicates. (*a–f*) Different migration rates, as shown in the title of each panel. The method producing each box plot appears below the *x* axis. In the "true" method, each mixture sample of 400 simulated fish is obtained by simulating additional leaves on the coalescent tree used to simulate the baseline samples. In addition to the bold line in the box plot, an extended broken line appears at the average $\hat{T}_{j,1}$ value obtained with the true method. Departures by the other methods from this line represent bias. In the other methods, each mixture sample of 400 fish is simulated by resampling from the baselines via either a parametric bootstrap method (PB-R, parametric bootstrap with baseline resampling; PB-NR, PB without baseline resampling) or a cross-validation method (CV-GC, cross-validation over gene copies; CV-SL, CV over single locus genotypes; CV-ML, CV over multilocus genotypes).



basis for two analyses to demonstrate practical consequences of using the conventional, biased method for predicting GSI accuracy. This baseline includes 22 231 fish from 166 Chinook salmon populations ranging from central California to Alaska (the Quinault Hatchery population was removed because of its small sample size). The baseline contains at least 144 fish from most of the 166 populations.

In the first analysis, we performed "100% mixture simulations" for each population in the baseline. In such a simulation, multiple mixture samples composed entirely of fish from a single population, say population *i*, are simulated and analyzed. The accuracy of GSI is assessed on the basis of how close the average value of $\widehat{\pi}_i$ is to 100%. Such 100% mixture simulations may not be a particularly realistic way of assessing the utility of a set of markers for GSI, but they have been used numerous times in the literature (for example, see Seeb and Crane (1999), Smith et al. (2005*b*), Habicht et al. (2007), etc.). We simulated one-hundred 100% mixtures of size $N = 200$ for each of the 166 populations in the baseline. These mixtures were simulated using both the CV-GC and the PB-R methods. For each method, we recorded the mean over the 100 samples of $\widehat{\pi}_i$. A separate set of simulations was performed with identical conditions except that each mixture was composed entirely of individuals from populations within a single one of 44 reporting groups that partition the 166 populations. Populations within each reporting group were assumed to contribute equally to the mixture. For the reporting group simulations, we recorded the mean over 100 replicates of the sum over populations *i* in the reporting group of $\widehat{\pi}_i$.

In the second analysis, we simulated 2500 mixed-fishery samples of size 400 fish using both the CV-GC and the PB-R methods. The value of $\boldsymbol{\pi}$ used to simulate the mixture samples was set close to the mixing proportions estimated using GSI from 735 Chinook salmon sampled from the Monterey Bay recreational fishery in the spring of 2006 (Table 1). Each sample was analyzed using the entire coastwide Chinook baseline. We recorded the mean and mean squared error (MSE) over all 2500 samples of $\widehat{\pi}_i$ for each population *i* represented in the mixture. Additionally, we assigned each fish in every simulated mixture to the population having highest posterior probability, approximated using the MLE, $\widehat{\pi}$. In other words, fish *j* was assigned to the population *i* that had the maximum value of

$$P_j = \frac{\widehat{\pi}_i T_{j,i}}{\sum_k \widehat{\pi}_k T_{j,k}}$$

We recorded the fraction of fish from each population assigned in this manner to the correct population.

Results

Simulation experiments

The results of our first simulation experiment demonstrate a profound bias in $\widehat{\boldsymbol{T}}_j$ when using the PB-NR and PB-R methods and show that the cross-validation methods provide essentially unbiased values of $\widehat{\boldsymbol{T}}_j$ (Fig. 2). Because this vector can be shown to be the sufficient statistic for estimating $\boldsymbol{\pi}$ using the conditional MLE method, bias in $\boldsymbol{T}_j$ indicates a serious problem for any simulation procedure. The bias is most pronounced when the two simulated populations are panmictic. In that case, the expected value of $\widehat{T}_{j,1}$ is, by a simple symmetry argument, 0.5, which is also the mean value obtained using the CV-ML, CV-SL, and CV-GC methods. In contrast, as expected for PB-NR, the distribution of $\widehat{T}_{j,1}$ is extremely biased, having a mean close to 0.7. Interestingly, using resampled baselines does little to correct the bias: the PB-R method shows an average value of $\widehat{T}_{j,1}$ of 0.66, far closer to the PB-NR value 0.7 than to the correct value of 0.5. As the migration rate decreases, the relative magnitude of the bias decreases, but the PB-NR and PB-R methods continue to deliver clearly biased values of $\widehat{T}_{j,1}$ all the way up to $4N_em = 10$ (expected $F_{ST} \approx 0.09$). All the cross-validation methods appear to remain unbiased at all migration rates.

The impact of this bias on the estimation of $\boldsymbol{\pi}$ is pronounced (Fig. 3). Even when a mixture is composed entirely

© 2008 NRC Canada

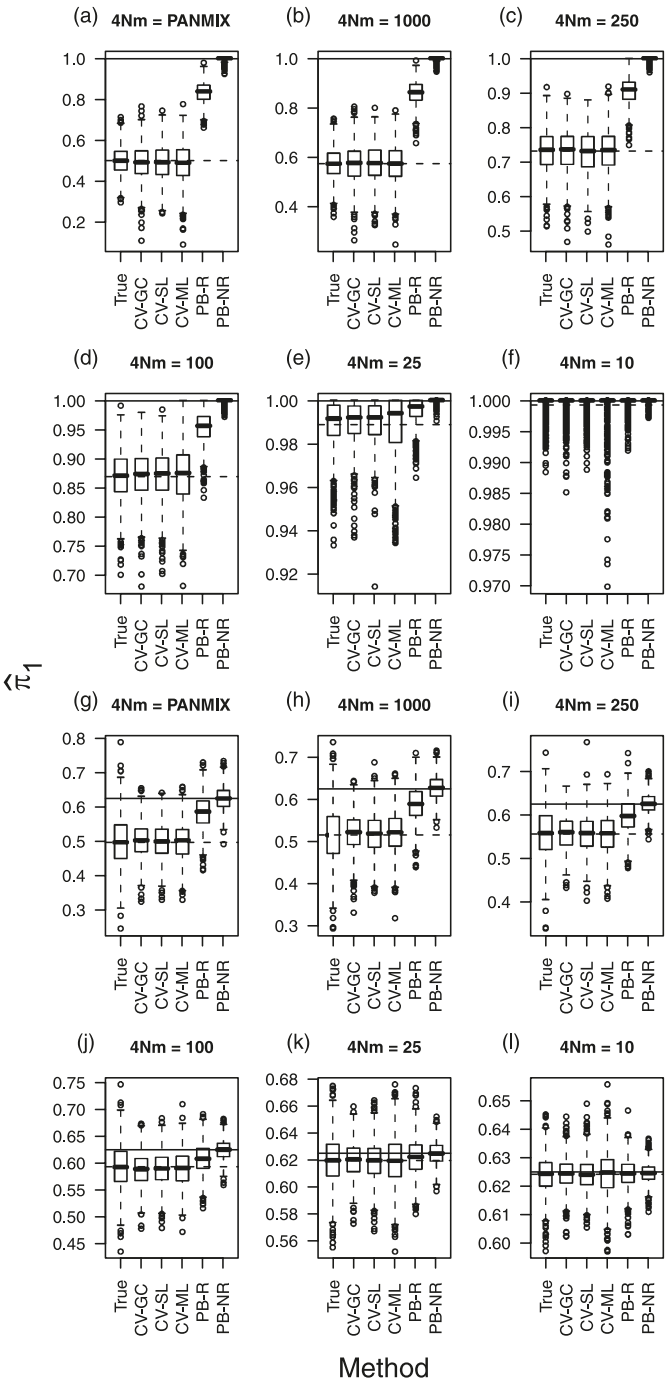


Fig. 3. Box plots of $\hat{\pi}_1$ in a two-population mixture model. Each box plot summarizes 1000 replicate simulations with $n_1 = n_2 = 144$ and $N = 400$. The actual proportion of fish from population 1 in each simulated mixture sample is represented by the horizontal, solid line (1.0 in panels *a–f* and 0.625 in panels *g–l*). The "true" box plot shows the true distribution of $\hat{\pi}_1$ obtained when the mixture sample is from the same population as the baseline sample but is taken separately from the baseline. The mean of this distribution is also denoted with a horizontal, broken line. The other box plots reflect mixtures obtained by resampling from the baselines. The distribution of $\hat{\pi}_1$ is notably biased for the PB-NR and PB-R methods. Such methods overestimate the accuracy of genetic stock identification (GSI). PB-R, parametric bootstrap with baseline resampling; PB-NR, PB without baseline resampling; CV-GC, cross-validation method over gene copies; CV-SL, CV over single locus genotypes; CV-ML, CV over multilocus genotypes.

of fish from population 1, if populations 1 and 2 are panmictic, then the true, expected value of $\hat{\pi}_1$ is 0.5. The average value of $\hat{\pi}_1$ obtained with the cross-validation methods is very close to 0.5. For the PB-R method, however, the average is 0.82. In this case, the bias inherent in $\hat{\boldsymbol{T}}_j$ using the PB-R method is magnified, creating an even more apparent bias in the estimates of $\hat{\pi}_1$. All the results in Fig. 3 show that the PB-R method provides biased estimates of π_1 at various migration rates and for two different true values of π_1, demonstrating that the conventional method overestimates, sometimes greatly so, the accuracy that can be expected from GSI given a particular baseline.

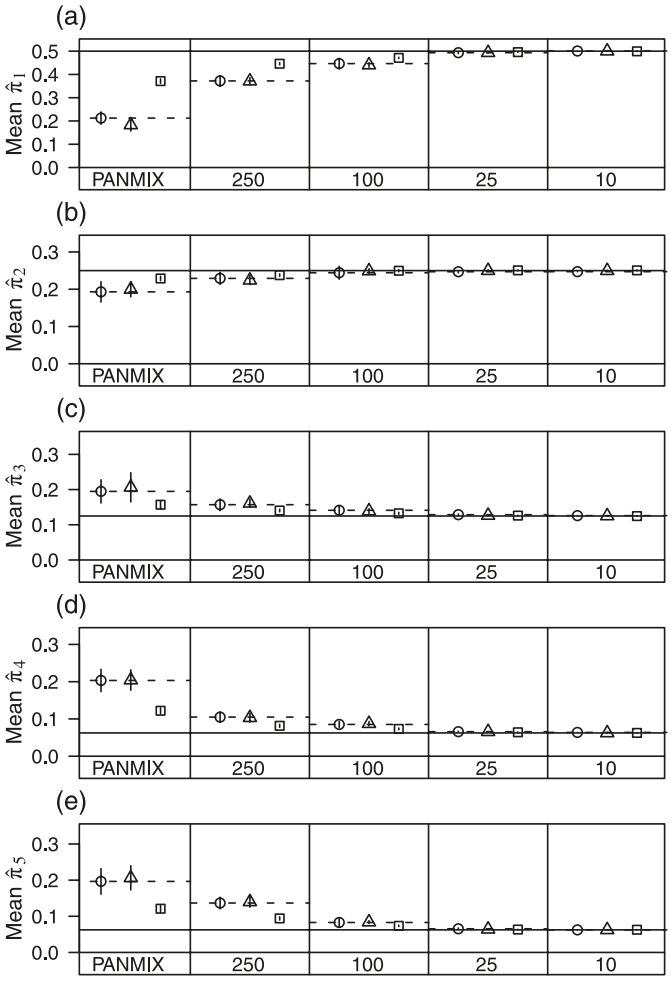
Though the mean values of $\hat{T}_{j,1}$ and $\hat{\pi}_1$ obtained using the cross-validation methods are very close to those obtained with the "true" method, it can be seen from the box plots that there are differences between the full distributions of $\hat{T}_{j,1}$ and $\hat{\pi}_1$ obtained by the different approaches. This is likely the case because $\hat{\boldsymbol{T}}_j$s simulated by the cross-validation methods are not conditionally independent given the baseline, as they are under the true model. Despite this nonindependence, however, it appears that the cross-validation methods provide an adequate approximation to the true distribution.

The five-population simulations deliver results (Fig. 4) similar to the two-population simulations. Under panmixia, the true expected value of $\hat{\pi}_i$ is 0.2 (one divided by the number of populations), and in fact, the observed mean $\hat{\pi}_i$ of the "true" simulations (circles in Fig. 4) do fall on 0.2 for the panmictic simulations. The mean $\hat{\pi}_i$ from the CV-GC method is also close to 0.2; the slight discrepancy, where it appears, is due, most likely, to the means being based on only 25 replicates. The PB-R method, however, tends to produce estimates of $\hat{\pi}_i$ that are much closer to the value of π_i used to simulate the mixtures. As before, when the divergence between populations increases, the bias becomes less apparent: with $4N_em = 25$, the average $\hat{\pi}_i$ is close to the true value of π_i for all methods. This suggests that the bias of the PB-R method should not be extremely problematic if all populations are well diverged genetically. These simulations, like the previous ones, confirm that the PB-R method yields consistently upward-biased estimates of GSI accuracy.

In our third simulation experiment, we found that the total number of alleles exerts a strong influence on the bias of the PB-R method (Fig. 5). The distribution of those alleles into a large or small number of loci had a secondary influence on the bias. For example, the degree of bias, measured in terms of mean $\hat{T}_{j,1}$, was nearly equal whether using 256 diallelic loci or 8 loci with 33 alleles each. (Because the frequency of one allele at each locus is completely determined by the remaining alleles, each diallelic locus has one independent allele and each locus with 33 alleles has 32 alleles, which, though not technically independent, are often referred to in the literature as "independent" alleles. Thus, 256 diallelic loci and 8 loci with 33 alleles have the same number of "independent" alleles.) Only when the number of alleles per locus approaches 64 does there seem to be a noticeable increase in bias resulting from the very low allele frequen-

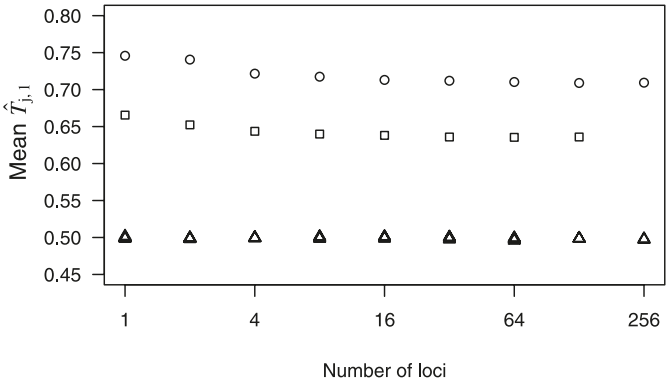
© 2008 NRC Canada

Fig. 4. Mean $\hat{\pi}$ in a simulation of five populations (*a*–*e*, populations 1–5, respectively). The true proportion that each population contributes to the mixture is shown by a horizontal, solid line. Results from simulations using different migration rates are designated by the migration rates listed at the bottom of each panel. Circles represent the mean of the average $\hat{\pi}$ expected if mixture samples are separate from baseline samples. These reflect the true accuracy that can be expected of genetic stock identification (GSI) given a baseline. Additionally, a horizontal, broken line is drawn at the level of this true expected $\hat{\pi}$. Triangles indicate the results for the CV-GC (cross-validation over gene copies) method. The open squares indicate the results for the PB-R (parametric bootstrap with baseline resampling) method. Vertical bars denote the standard deviation of the means. The CV-GC values correspond closely to the true, expected accuracy of GSI. The PB-R values are biased toward the true value of π, suggesting greater accuracy than is truly available from a given baseline. The differences between the methods are minimal at low migration rates (high genetic divergence) but are remarkable at high migration rates (low divergence).



cies, and even this increase is minor. The results also show that with more data (i.e., more alleles in total), the bias increases. In other words, as more and more genetic markers become available and are applied to resolve mixtures of closely related populations, it becomes increasingly vital that an unbiased method of GSI accuracy prediction be used. Regardless of the number of alleles or loci in this simulation, all the cross-validation methods yielded unbiased

Fig. 5. Effect of the number of alleles per locus on the magnitude of the bias. Total number of alleles is constant, but number of loci changes. Cross-validation methods (CV-GC, CV over gene copies; CV-SL, CV over single locus genotypes; CV-ML, CV over multilocus genotypes) with 128 and 256 alleles are all denoted by open triangles. Parametric bootstrap method with baseline resampling (PB-R) with 128 and 256 alleles are denoted by open squares and open circles, respectively. In a two-population scenario, the amount of bias in the PB-R method is shown here to be roughly constant, whether alleles are apportioned into just a few loci, each with many alleles, or into many loci, each with only two alleles.



values of $\widehat{T}_{j,1}$. (Recall that, by symmetry, the true expectation of $\widehat{T}_{j,1}$ is 0.5.)

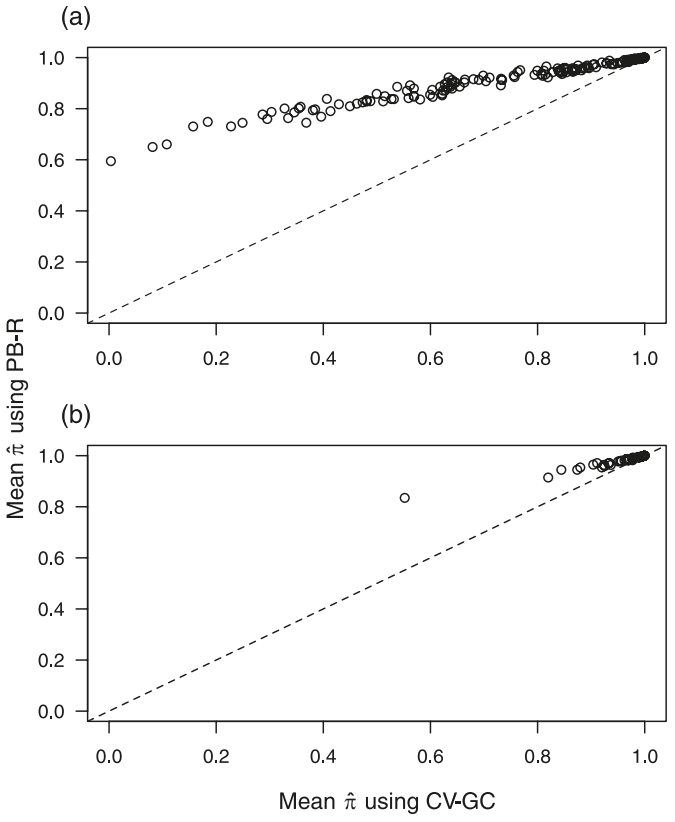
Analysis of empirical data

The simulations described above established that the cross-validation methods provide largely unbiased estimates of GSI accuracy. Now, the results obtained using the CV-GC and PB-R methods applied to real data give us a sometimes alarming picture of the practical implications of the PB-R method's bias. This is plainly apparent in the 100% simulated mixtures of populations in the GAPS baseline (Fig. 6*a*). Using the PB-R method, only 17 of 166 populations had a mean $\widehat{\pi}_i < 0.8$. By contrast, the CV-GC method reveals that fully 78 of 166 populations have mean $\widehat{\pi}_i$ values less than 0.8. Many populations having mean $\widehat{\pi}_i \approx 0.85$ under the PB-R method have mean $\widehat{\pi}_i$ close to 0.6 using the CV-GC method. For all those populations, the predicted accuracy using PB-R is biased upward by some 40% from its unbiased value obtained using CV-GC. When focusing on reporting units, the bias appears to be not as great because, overall, accuracy is higher for estimating contributions of reporting units than for estimating contributions of their consituent populations. Nonetheless, there is some bias apparent (Fig. 6*b*). When using PB-R, none of the 44 reporting units (except for one, composed of a single population) has mean π values < 0.9, whereas when using the CV-GC method, 4 of them do.

The fact that the bias of the PB-R method is of less consequence when genetic differentiation is substantial between populations was confirmed in our simulations mimicking the Monterey Bay recreational fishery (Table 1). There were no clear, systematic differences in the mean estimates of $\boldsymbol{\pi}$ between the CV-GC and the PB-R methods relative to the value of $\boldsymbol{\pi}$ used to drive the simulation. In this case, the lack of apparent bias in PB-R is not unexpected; even when

© 2008 NRC Canada

Fig. 6. One-hundred percent mixture results. (*a*) Each circle represents a single one of 166 populations in the coastwide Chinook baseline. (*b*) Each circle represents one of 44 regions (or reporting units) into which the 166 populations fall. The position along the *x* axis indicates the mean over 100 simulations of the proportion of that population (or region) estimated from a mixture sample composed exclusively of members from that single population (or region) simulated using the CV-GC method (cross validation over gene copies). The *y* axis indicates the results using the parametric bootstrap method with baseline resampling (PB-R). The broken line falls along $y = x$. The PB-R method is biased upward relative to the CV-GC method, overestimating genetic stock identification (GSI) accuracy (greatly so in the case of some individual populations).



using the CV-GC method, the mean $\widehat{\pi}_i$ is close to the true value of π_i for all the populations. If you are getting an excellent estimate from an unbiased method already, then the bias of the PB-R method cannot contribute much spurious, additional accuracy. Additionally, the MSE of the estimates were not substantially different between the two methods. This, also, is to be expected as most of the MSE in this case results from the multinomial variance associated with the number of fish from each population appearing in each simulated mixture sample, and not from the inaccuracy of genetic discrimination between the populations.

Though there was no discernible difference between using CV-GC and PB-R in the estimation of $\widehat{\boldsymbol{\pi}}$ in the Monterey Bay Fishery simulations, there was a marked difference in the observed rate of correct assignment of individual fish to their population of origin (see Individual assignment columns in Table 1). The fraction of correctly assigned fish using the PB-R method was always biased high relative to the fraction using the CV-GC method. For some populations, this bias was very pronounced. For example, the estimated rate of correct allocation of salmon from Stanislaus River and Butte Creek Fall, using PB-R, is over three times too high. For applications requiring a careful assessment of the accuracy of assignment of individual fish from mixed fisheries, it is critically important to use an unbiased method like one of the cross-validation methods in preference to the PB-R method.

Discussion

In this paper, we have demonstrated that the conventional simulation method, which we call the PB-R method, invariably provides a biased, overly optimistic prediction of the accuracy achievable with genetic stock identification. The PB-R simulation method involves resampling from the observed baseline to produce a simulated mixture and a simulated baseline that is used to analyze the simulated mixture. It has been widely appreciated that the PB-NR method, analyzing the simulated mixture with the observed baseline, might yield biased results, and we have shown this to be the case. However, for over 20 years, it seems to have been commonly believed that resampling the baseline as in the PB-R method will minimize or eliminate that bias; it is clear from our computer simulations that it does not.

We have also described and implemented a method that provides essentially unbiased esimates of GSI accuracy. The three versions of this method (CV-ML, CV-SL, and CV-GC) are all based on the leave-one-out, cross-validation procedure. Using simulated data, we first confirmed that our cross-validation methods give essentially unbiased results. Then, by applying the PB-R and CV-GC methods to empirical data sets, we were able to assess the practical consequences of the PB-R method's bias.

We wish to stress that our method yields essentially unbiased estimates of GSI accuracy only within the confines of the asssumptions of the model, and particularly the assumption that all populations in the mixture are accurately represented in the baseline. Clearly, if individuals will appear in the mixture from populations that do not occur in the baseline, or if there is temporal or spatial structure in the baseline populations that is not well represented by the baseline sample, then not even the CV-GC, CV-SL, or CV-ML methods will reflect the accuracy that you can expect in performing GSI. It would seem that there is no way to systematically and correctly account for the possibility of individuals from unsampled populations in the mixture.

There is little apparent difference in results for the three cross-validation methods that we tried, but these differences might be larger under certain conditions. For example, CV-ML has the advantage that if some loci (or pairs of loci) are out of Hardy–Weinberg (or linkage) equilibrium, that will be reflected in the simulated genotypes; in addition, this option might best reflect the distribution of missing data. On the other hand, CV-SL might be preferable if individual loci show Hardy–Weinberg departures but pairs of loci are in linkage equilibrium. Finally, CV-GC is perhaps the most flexible option, as it can be used when only allele frequencies (and not genotypes) are available, and the space of possible simulated genotypes is largest with that option.

To facilitate analysis of the large number of simulated

© 2008 NRC Canada

data sets, we only evaluated results for the conditional GSI method. Although the unconditional GSI and Bayesian methods have some potential advantages for genetic mixture analysis, as currently implemented, they are not designed to deal with the problem of inflated F'_{ST} in simulated mixtures due to baseline sampling error discussed in this paper. Therefore, we would expect the same general patterns of bias from these methods as we found for conditional GSI.

The impact of the bias is greatest in situations involving closely related populations, and the extent of bias increases as more loci and alleles are added to the data set. Both of these observations make sense intuitively based on the concepts of signal-to-noise ratio and statistical power. In closely related populations, the noise (spurious divergence due to sampling error) is large relative to the true population divergence (signal), and this exacerbates the bias. For any given inflated value of F'_{ST}, the apparent (but inflated) power to resolve the mixture increases as more loci and more alleles are used. Further simulations (not shown) show that this latter effect occurs whether the genetic markers in question are microsatellites or SNPs.

In the fisheries literature, populations yielding an average estimated mixture proportion of 90% from 100% mixture simulations are often considered highly identifiable in mixed fisheries (Smith et al. 2005*b*). Using the essentially unbiased CV-GC method, only 54 of the 166 populations in the GAPS data set would be considered highly identifiable. Using the traditional PB-R simulations, however, 110 populations in the GAPS data set exceeded this highly identifiable threshold. Thus, in the case of the GAPS data, the PB-R method has misclassified more than half of the 110 populations placed in the highly identifiable category. We expect that most fisheries managers will find this degree of bias in the PB-R method unacceptable. One-hundred percent mixture simulations, using the PB-R method and typically implemented in SPAM (Alaska Department of Fish and Game 2000) or GMA (Kalinowski 2003), have been employed in a great number of studies. Unfortunately, our work indicates that re-evaluation of those past results may be warranted.

When inference for individual populations is not critical and similar populations can be aggregated into larger reporting units (collections of multiple populations), the consequences of the PB-R method's bias are not so dramatic. This occurs because the genetic divergence between members of different reporting units often is large enough that the additional bias from using the PB-R method is relatively less noticeable. Our results show that although there is still a consistent bias in the 100% mixture results for 44 reporting units, the relative magnitude of the bias is not nearly so great as with individual populations.

When multiple populations contribute to a mixture, the effects of the PB-R method's bias can be more difficult to predict. This is apparent from our simulated Monterey Bay Recreational Fishery samples. In this case, the PB-R and the CV-GC methods delivered similar estimates of the accuracy with which mixing proportions of individual populations can be estimated (i.e., the means and the mean-squared errors from both methods were roughly comparable). This is likely because the actual numbers of fish from each population in these simulated mixture samples were themselves random variables (with means equal to N times the true $\boldsymbol{\pi}$), and the variance associated with each one resulting from this random variation may have been large enough to obscure the differences between the PB-R and the CV-GC methods. On the other hand, if one intends to use genetic data not only for estimating $\boldsymbol{\pi}$, but also for allocating individual fish to populations, then the PB-R method still provides strongly biased results. In the Monterey Bay example, when individuals were assigned to the population having highest posterior probability, the rate of correct assignment predicted by the PB-R method was, for some populations, over three times greater than the unbiased prediction from the CV-GC method.

Our results apply to simulations in which hypothetical mixture samples are simulated using the baseline as input. When in possession of both a baseline sample and a separate sample from a mixed fishery, the sampling distribution of $\widehat{\boldsymbol{\pi}}$ is sometimes estimated by repeatedly bootstrap resampling a new mixture sample from the real mixture sample and a new baseline sample from the observed baseline sample and then estimating $\boldsymbol{\pi}$ from the bootstrapped samples (Fournier et al. 1984). Because this procedure does not induce a positive correlation (with respect to the true population allele frequencies) between the bootstrapped mixtures and the baselines, it should not suffer from the same degree of bias as the PB-R method. In fact, this is apparent by comparing the OBSERVED, SIMULATE, and BOOTSTRAP columns of tables 10.2 and 10.3 in Pella and Milner (1987, p. 254). It should be kept in mind, however, that such a bootstrapping scheme yields only the sampling distribution of $\widehat{\boldsymbol{\pi}}$, which may not reflect the range of possible true values of $\boldsymbol{\pi}$ that could yield such estimates, especially when closely related populations are present in the baseline.

We introduced this problem in terms of the biased value of F_{ST} between populations obtained when the effect of sampling is not taken into account. If one desires merely an unbiased estimator of F_{ST}, then such an estimator is available using an analysis of variance approach (Weir and Cockerham 1984). On the basis of such an unbiased estimate of F_{ST}, it would be possible to simulate genes from two populations, adjusting the results so as to achieve samples having properties that would be expected of two populations diverged by an amount F_{ST} rather than by F'_{ST} (see Introduction). Such maneuvers might allow for simulations of mixtures and baselines that would provide a less biased (than PB-R) prediction of GSI accuracy, using an approach that is of an entirely different character than ours. In fact, such a method was used to parametrize simulations of spring- and fall-run Chinook salmon in the Trinity River (Kinziger et al. 2008) designed to help interpret the results of the program *structure* (Pritchard et al. 2000). Though such an approach is useful when only two populations are involved, its implementation becomes more complicated when the baseline includes a large number of populations at widely varying degrees of divergence. A related approach would involve using estimates of baseline frequencies that were adjusted by some shrinkage toward a grand mean allele frequency. For predicting GSI accuracy, it seems unlikely that such shrinkage methods would be as elegant and simple to implement as the cross-validation method; however, it could prove very useful in some contexts.

The bias effect we have described might not be limited

© 2008 NRC Canada

merely to genetic stock identification in fisheries; it could also confound other accuracy estimation problems in molecular ecology. One example occurs with Bayesian, model-based clustering approaches like those implemented in *structure* (Pritchard et al. 2000), *NewHybrids* (Anderson and Thompson 2002), and *BayesAss+* (Wilson and Rannala 2003). With closely related populations, it may be difficult to interpret the output from these programs, and it is thus becoming increasingly common to run the programs multiple times on data simulated to look like the original data set and, from the ensemble of outputs, arrive at an interpretation of the original results. For example, Nielsen et al. (2003) take this approach using the program *structure* to argue that there is an Atlantic cod hybrid zone. We believe that doing this type of simulation is an important step in applying such Bayesian clustering methods; however, our current work demonstrates that particular care must be taken in choosing how to perform the simulations. For example, naively simulating new data sets from the posterior mean allele frequencies in the clusters suffers from the same problems as the PB-R method and could lead one to conclude that more power is available for resolving clusters than really exists. It might be possible, however, to design such simulations using the same principles of leave-one-out cross validation as applied here. We are currently investigating the potential for this in the context of Bayesian model checking and sensitivity analysis executed during the run time of the Markov chain Monte Carlo algorithm of each program.

We foresee that more and more genetic markers will become available, and the demands placed on them for fisheries management will only increase. Managers will likely seek resolution between increasingly closely related populations, and applications for allocation of individual fish will continue to develop in fisheries. In such a climate, it is important to have accurate methods for assessing statistical power for GSI. We suggest that the software packages currently in use be updated to include, as the default option, at least one of the cross-validation methods described here. ONCOR, a Windows-based program, for implementing the simulations described here is available at http://www.montana.edu/kalinowski. Another program, *gsi_sim*, with a command line interface suitable for Unix-like operating systems is available at http://swfsc.noaa.gov/staff.aspx?id=740.

Acknowledgments

This work grew out of concerns expressed to us by Paul Moran and David Teel and benefited from discussions with Ken Warheit, Carlos Garza, and Devon Pearse. We thank Michele Masuda and an anonymous reviewer for useful comments on this paper and Anton Antonovich for clarifying the inner workings of SPAM and for performing simulations to confirm that *gsi_sim* with the PB-R option and SPAM yield comparable results.

References

Alaska Department of Fish and Game. 2000. SPAM, version 3.2. User's guide. Technical Report, Alaska Department of Fish and Game, Commercial Fisheries Division, Gene Conservation Lab, 333 Raspberry Road, Anchorage, AK 99518, USA. Special Publication No. 15.

Anderson, E.C., and Thompson, E.A. 2002. A model-based method for identifying species hybrids using multilocus genetic data. Genetics, **160**: 1217–1229. PMID:11901135.

Beacham, T.D., Candy, J.R., Jonsen, K.L., Supernault, K.J., Wetklo, M., Deng, L., Miller, K.M., and Withler, R.E. 2006. Estimation of stock composition and individual identication of chinook salmon across the Pacific rim by use of microsatellite variation. Trans. Am. Fish. Soc. **135**: 861–888. doi:10.1577/T05-241.1.

Casella, G., and Berger, R.L. 1990. Statistical inference. Duxbury Press, Belmont, Calif.

Chakraborty, R., and Leimar, O. 1987. Genetic variation within a subdivided population. *In* Population genetics and fishery management. *Edited by* N. Ryman and F. Utter. University of Washington Press, Seattle, Wash. pp. 89–102.

Fournier, D.A., Beacham, T.D., Riddell, B.E., and Busack, C.A. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. **41**: 400–408. doi:10.1139/f84-047.

Grant, W., Milner, G., Krasnowski, P., and Utter, F. 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. doi:10.1139/f80-159.

Habicht, C., Seeb, L.W., and Seeb, J. 2007. Genetic and ecological divergence defines population structure of sockeye salmon populations returning to Bristol Bay, Alaska, and provides a tool for admixture analysis. Trans. Am. Fish. Soc. **136**: 82–94. doi:10.1577/T06-001.1.

Hudson, R.R. 1990. Gene genealogies and the coalescent process. *In* Oxford surveys in evolutionary biology. *Edited by* D. Futuyma and J. Antonovics. Oxford University Press, Oxford, UK. pp. 1–44.

Hudson, R.R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics, **18**: 337–338. doi:10.1093/bioinformatics/18.2.337. PMID:11847089.

Kalinowski, S.T. 2003. Genetic mixture analysis 1.0. Department of Ecology, Montana State University, Bozeman, MT 59717, USA. Technical Report. Available online: http://www.montana.edu/kalinowski.

Kalinowski, S.T. 2004. Genetic polymorphism and mixed-stock fisheries analysis. Can. J. Fish. Aquat. Sci. **61**: 1075–1082. doi:10.1139/f04-060.

Kinziger, A.P., Loudenslager, E.J., Hankin, D.G., Anderson, E.C., and Garza, J.C. 2008. Hybridization between spring-run and fall-run chinook salmon returning to Trinity River, CA. North Am. J. Fish. Manage. In press.

Koljonen, M.J., Pella, J., and Masuda, M. 2005. Classical individual assignments versus mixture modeling to estimate stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite data. Can. J. Fish. Aquat. Sci. **62**: 2143–2158. doi:10.1139/f05-128.

McParland, T.L., Ferguson, M.M., and Liskauskas, A.P. 1999. Genetic population structure and mixed-stock analysis of walleyes in the Lake Erie – Lake Huron corridor using allozyme and mitochondrial DNA markers. Trans. Am. Fish. Soc. **128**: 1055–1067. doi:10.1577/1548-8659(1999)128<1055:GPSAMS>2.0.CO;2.

Millar, R.B. 1987. Maximum likelihood estimation of mixed stock fishery composition. Can. J. Fish. Aquat. Sci. **44**: 583–590. doi:10.1139/f87-071.

Milner, G.B., Teel, D.J., Utter, F.M., and Burley, C.L. 1981. National Marine Fisheries Service, Columbia River stock identification study: validation of genetic method. Final Report to Bonneville Power Administration (Contract No. 1980BP18488), BPA Report DOE/BP-18488-1.

© 2008 NRC Canada

Neaves, P.I., Wallace, C.G., Candy, J.R., and Beacham, T.D. 2005. CBayes: computer program for mixed stock analysis of allelic data. Version 3.0. Free program available from the authors online: http://www.pac.dfo-mpo.gc.ca/sci/mgl/Cbayes e.htm.

Nielsen, E.E., Hansen, M.M., Ruzzante, D.E., Meldrup, D., and Gronkjaer, P. 2003. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. Mol. Ecol. **12**: 1497–1508. doi:10.1046/j.1365-294X.2003.01819.x. PMID:12755878.

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

Pella, J., and Masuda, M. 2006. The Gibbs and split–merge sampler for population mixture analysis from genetic data with incomplete baselines. Can. J. Fish. Aquat. Sci. **63**: 576–596. doi:10.1139/f05-224.

Pella, J., and Milner, G.B. 1987. Use of genetic marks in stock composition analysis. *In* Genetics and fishery management. *Edited by* N. Ryman and F. Utter. University of Washington Press, Seattle, Wash. pp. 247–276.

Piry, S., Alapetite, A., Cornuet, J.M., Paetkau, D., Baudouin, L., and Estoup, A. 2004. GeneClass2: a software for genetic assignment and first-generation migrant detection. J. Hered. **95**: 536–539. doi:10.1093/jhered/esh074. PMID:15475402.

Pritchard, J.K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. Genetics, **155**: 945–959. PMID:10835412.

Rannala, B., and Mountain, J.L. 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. U.S.A. **94**: 9197–9201. doi:10.1073/pnas.94.17.9197. PMID:9256459.

Seeb, L.W., Antonovich, A., Banks, A.A., Beacham, T.D., Bellinger, A.R., Blankenship, S.M., Campbell, A.R., Decovich, N.A., Garza, J.C., Guthrie, C.M., Lundrigan, T.A., Moran, P., Narum, S.R., Stephenson, J.J., Supernault, K.J., Teel, D.J., Templin, W.D., Wenburg, J.K., Young, S.E., and Smith, C.T. 2007. Development of a standardized DNA database for Chinook salmon. Fisheries, **32**: 540–552.

Seeb, L.W., and Crane, P.A. 1999. Allozymes and mitochondrial DNA discriminate Asian and North American populations of chum salmon in mixed-stock fisheries along the south coast of the Alaska Peninsula. Trans. Am. Fish. Soc. **128**: 88–103. doi:10.1577/1548-8659(1999)128<0088:AAMDDA>2.0.CO;2.

Smith, C.T., Elfstrom, C.M., Seeb, L.W., and Seeb, J.E. 2005*a*. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol. Ecol. **14**: 4193–4203. doi:10.1111/j.1365-294X.2005.02731.x. PMID:16262869.

Smith, C.T., Templin, W.D., Seeb, J.E., and Seeb, L.W. 2005*b*. Single nucleotide polymorphisms provide accurate estimates of the proportions of U.S. and Canadian chinook salmon caught in Yukon River fisheries. N. Am. J. Fish. Manag. **25**: 944–953. doi:10.1577/M04-143.1.

Smouse, P.E., Waples, R.S., and Tworek, J.A. 1990. A genetic mixture analysis for use with incomplete source population data. Can. J. Fish. Aquat. Sci. **47**: 620–634. doi:10.1139/f90-070.

Spielman, R.S., and Smouse, P.E. 1976. Multivariate classification of human populations. I. Allocation of Yanomama Indians to villages. Am. J. Hum. Genet. **28**: 317–331. PMID:821344.

Waldman, J.R., Richards, R.A., Schill, W.B., Wirgin, I., and Fabrizio, M.C. 1997. An empirical comparison of stock identification techniques applied to striped bass. Trans. Am. Fish. Soc. **126**: 369–385. doi:10.1577/1548-8659(1997)126<0369:AECOSI>2.3.CO;2.

Weir, B.S., and Cockerham, C.C. 1984. Estimating F-statistics for the analysis of population structure. Evolution, **38**: 1358–1370. doi:10.2307/2408641.

Wilson, G.A., and Rannala, B. 2003. Bayesian inference of recent migration rates using multilocus genotypes. Genetics, **163**: 1177–1191. PMID:12663554.

Winans, G.A., Paquin, M.M., Van Doornik, D.M., Baker, B.M., Thornton, P., Rawding, D., Marshall, A., Moran, P., and Kalinowski, S. 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. doi:10.1577/M03-052.1.

Wood, C.C., McKinnell, S., Mulligan, T.J., and Fournier, D.A. 1987. Stock identification with the maximum-likelihood mixture model: sensitivity analysis and application to complex problems. Can. J. Fish. Aquat. Sci. **44**: 866–881. doi:10.1139/f87-105.

Wright, S. 1978. Evolution and the genetics of populations. Vol. IV. Variability within and among natural populations. University of Chicago Press, Chicago, Ill.

© 2008 NRC Canada