Comparison of the accuracy and consistency of likelihood-based estimation routines for genetic stock identification

Travis O. Brenden1,2*, James R. Bence1,2, Weihai Liu1,2, Iyob Tsehaye1,2 and Kim T. Scribner1

1Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824, USA; and 2Quantitative Fisheries Center, Michigan State University, East Lansing, MI 48824, USA

Summary

1. Genetic stock identification (GSI) frequently is used to assess spawning/breeding population contributions to mixtures of individuals. Although multiple estimation routines are available for conducting GSI, their performance may vary depending on characteristics of assessed source populations and mixtures and of employed genetic markers.

2. We conducted simulations to compare performance of several likelihood-based GSI estimation routines. Estimation routines were implemented in SPAM, ONCOR and AD Model Builder (ADMB). Two ADMB routines were evaluated, one based on conditional maximum likelihood estimation (ADMB-MLE), similar to SPAM and ONCOR, and one based on conditional penalized maximum likelihood estimation (ADMB-PMLE). The simulations examined how performance varied by number of source populations, population divergence levels, number of evaluated loci, and source population and mixture sample sizes. Evaluations included scenarios with many loci with low levels of polymorphism for assessing performance when single nucleotide polymorphism (SNP) markers are incorporated in analyses.

3. Mixture sample size and source population genetic divergence accounted for most of the explained variability in simulation results. Overall, routines based on conditional maximum likelihood estimation (SPAM, ONCOR and ADMB-MLE) had similar levels of accuracy, including scenarios mimicking SNP markers, with SPAM having slightly better accuracy than ONCOR and ADMB-MLE. The accuracy of the ADMB-PMLE routine in many scenarios was noticeably poorer than the other routines, although in some instances accuracy of the ADMB-PMLE estimates approached the other routines with large mixture sample sizes. SPAM, ONCOR and ADMB-MLE also generally had similar levels of performance with respect to consistency, whereas ADMB-PMLE varied widely in consistency due in part to poor accuracy.

4. Because SPAM and ONCOR typically performed better than the ADMB-MLE routine, there appears to be little need for users to program their own likelihood-based estimation routines for standard GSI analyses, although for specialized applications (e.g. modelling contributions as functions of ecological or demographic features), it may be necessary for users to program their own routines. Given the performance of the ADMB-PMLE routine, additional research is needed to determine an appropriate configuration (e.g. penalty, optimization algorithm) for a penalized maximum likelihood GSI estimator.

Key-words: AD Model Builder, maximum likelihood, mixed stock analysis, ONCOR, penalized maximum likelihood, SPAM, stochastic simulations

Introduction

Genetic stock identification (GSI), which is a form of mixed stock analysis, is commonly used to estimate contributions of spawning/breeding populations to mixtures of individuals. GSI has been applied to both terrestrial (e.g. Scribner et al. 2003) and aquatic (e.g. reptiles: Bolker et al. 2003; Prosdocimi et al. 2014; mammals: Baker et al. 2000; Lukoschek et al. 2009; fishes: Ruzzante et al. 2000; Beacham, Jonsen & Wallace 2012; Ensing et al. 2013) species and in several contexts including estimating composition of harvest from recreational or commercial fisheries (Pella & Milner 1987; Begg, Friedland & Pearce 1999) and forensic identification of possible illegally harvested individuals (Ogden 2008; Chapman, Pinhal & Shivji 2009; Lukoschek et al. 2009). Use of GSI has become common due to availability of highly polymorphic and high-throughput DNA markers, which has made it possible to distinguish considerable genetic variability among source populations for many species (Hauser & Seeb 2008).

Both model-based and classification-type approaches have been used for GSI analyses (Pella & Milner 1987; Millar 1990).
Model-based approaches are based on finite-mixture distributions with inferences about population contributions made using either likelihood- or Bayesian-based methods. Model-based approaches have several benefits, including the ability to assign measures of confidence to contribution estimates, model contributions as functions of ecological or demographic parameters and incorporate prior knowledge as to population contributions (Pella & Masuda 2001; Guo, Dey & Holsinger 2008). Several software programs are available for conducting model-based GSI, including BAYES (Pella & Masuda 2001), the Statistical Package for the Analysis of Mixtures (SPAM) (ADFG 2003, eBAYES (Neaves et al. 2005) and ONCOR (Kalinowski, Manlove & Taper 2007). Additionally, any numerical or statistical software package with optimization capabilities can be used to develop GSI estimation routines.

Because software packages may differ in optimization algorithms, termination criteria, or use special features to aid in parameter estimation, differences in accuracy among estimation routines are possible, with some perhaps performing better than others in particular circumstances. Although there have been some empirical comparisons of estimation routines (Koljonen, Pella & Masuda 2005; Ensing et al. 2013), comparisons using simulated or known-origin data have not been widely conducted. The advantage of simulated or known-origin data for such comparisons is that actual contributions will be known, so accuracy of estimation routines can truly be evaluated. Araujo et al. (2014) used simulated and known-origin data to compare performance of two GSI estimation routines: cBAYES (a Bayesian-based estimation routine) and ONCOR (a likelihood-based estimation routine). ONCOR was more accurate when genetic divergence among source populations was low, but at higher divergences, there was little difference between the routines (Araujo et al. 2014). Beacham et al. (2005) compared performance of cBAYES and SPAM (a likelihood-based estimation routine) using known-origin and 100% mixture simulations (i.e. simulated mixtures comprised entirely of individuals from a single population) and found that cBAYES was marginally more accurate for 100% mixture simulations, while SPAM was marginally more accurate for known-origin mixtures when source populations were treated individually (i.e. not pooled into groups).

Although the Beacham et al. (2005) and Araujo et al. (2014) comparisons helped shed light on relative performance of some available GSI estimation routines, there is still much to be learned about performance of different routines under various conditions, such as when analyses incorporate single nucleotide polymorphisms (SNPs). SNPs generally have two alleles per locus, but markers can be identified quickly and efficiently, so the number of loci that can be incorporated in analyses is greater than with other markers. Consequently, the use of SNPs in GSI analyses has become common (Ackerman, Habicht & Seeb 2011; Bradbury et al. 2011; Beacham, Jonsen & Wallace 2012; Clemento et al. 2014; Larson et al. 2014). However, given their biallelic features as well as the potential for more loci to be incorporated in GSI analyses, there is the possibility for SNPs to affect performance of estimation routines.

The purpose of this research was to use simulations to evaluate performance of several likelihood-based GSI estimation routines. Part of our reason for focusing on likelihood-based estimation routines was the findings of Beacham et al. (2005) and Araujo et al. (2014) that likelihood-based estimation routines performed as well as, or better, than Bayesian-based estimation routines in situations resembling typical GSI analyses (i.e. mixtures consisting of individuals from multiple populations with some populations proportionally contributing fewer individuals). Further, given the total number of simulations conducted for this research (270 thousand), using Bayesian-based estimation routines that relied on Markov chain Monte Carlo (MCMC) methods to characterize posterior probability distributions of estimated parameters was not feasible in a timely fashion. Evaluated estimation routines included those implemented in SPAM and ONCOR, which are based on conditional maximum likelihood estimation. Two estimation routines that we programmed in AD Model Builder (Fournier et al. 2012) were also evaluated. AD Model Builder is an open-source statistical modelling package that uses automatic differentiation and produces accurate and stable parameter estimates even for heavily parameterized models (Fournier et al. 2012). Because of these features, there might be benefits to using the software for GSI analyses. Additionally, the programming of one’s own estimation routine in a software package such as AD Model Builder could allow for more customized GSI applications (e.g. estimating contributions as functions of demographic or environmental features). Consequently, we were interested in comparing performance among existing software packages and user-developed estimation routines. One of the AD Model Builder routines was based on conditional maximum likelihood estimation, while the other was based on conditional penalized maximum likelihood estimation. Penalized maximum likelihood estimation is also referred to as the highest posterior density estimation due to point estimates equating to modes of the posterior probability distributions for the parameters (i.e. penalized maximum likelihood estimation has a Bayesian interpretation but does not use MCMC algorithms to characterize posterior probability distributions). Our aim was to provide information to those interested in conducting GSI analyses about the performance of different likelihood-based estimation routines when applied across a range of possible circumstances.

Materials and methods

MIXTURE AND POPULATION SIMULATIONS

We followed the methodology of Guo, Dey & Holsinger (2008) for simulating source population and mixture data, and programmed it in AD Model Builder. A common hyperpopulation $\psi$ of fixed allele frequencies for evaluated loci was assumed for the source populations. For each iteration, this hyperpopulation was generated by $L_L (L = \text{total number of evaluated loci})$ random draws from Dirichlet distributions with concentration parameters (total number of concentration parameters equal to the total number of alleles per locus) assumed to be equal to 1. The number of alleles per locus depended on the number of
evaluated loci and for some scenarios was randomly determined (see Simulations and Evaluated Factors). Actual allele frequencies for each locus for each of the source populations differed from the hyperpopulation based on the assumed divergence factor $\theta$, which as noted by Guo, Dey & Holsinger (2008) is analogous to Wright’s $F_{ST}$. The actual allele frequencies for locus $i$ for the source populations for each simulation were generated by $I(t = \text{total number of populations})$ random draws from Dirichlet distributions with concentration parameters equal to $((1 - 0)/0)\psi_i$. Each population at each locus was assumed to be in Hardy–Weinberg equilibrium.

In actual GSI analyses, allele relative frequencies for source populations are estimated by genotyping tissue samples from individuals collected from them. Consequently, observation error can result in differences between assessed and true allele frequencies. We accounted for this error by assuming that sample genotypes from which allele relative frequencies were calculated were drawn randomly from multinomial distributions, with probabilities equal to the expected genotype frequencies and the number of trials equal to the population sample size under evaluation (see Simulations and Evaluated Factors). These ‘observed’ genotypes were then used to calculate allele relative frequencies for the source populations.

Data from the mixture were also assumed to be influenced by observation error. First, the number of sampled individuals in the mixture that came from each source population was determined by random draw from a multinomial distribution with probabilities and number of trials equal to the assumed population contributions and mixture sample size under evaluation (see Simulations and Evaluated Factors). Then, the genotypes of individuals in the mixture that came from each of the source populations were generated by random draws from multinomial distributions with probabilities equal to the expected genotype frequencies for the source populations and the number of trials equal to the number of individuals in the mixture that came from the source populations.

ESTIMATION ROUTINES

Description of the AD Model Builder estimation routines developed for this research is provided in Appendix S1 (Supporting information). As previously indicated, there were two AD Model Builder routines, one based on conditional maximum likelihood estimation (hereafter ADMB-ML) and the second by conditional penalized maximum likelihood estimation (hereafter ADMB-PMLE). The primary difference between these routines was that a ‘penalty’ was added to the ADMB-PMLE objective function for source population contributions deviating from prior assumptions as to their possible values. In penalized maximum likelihood estimation, the ‘penalty’ can help stabilize parameter estimates by regularizing the numerical search routine. We used a Dirichlet probability density function with concentration parameters equal to the inverse of the number of source populations as a vague prior to the source population contribution estimates as the basis for the ‘penalty’. Such a prior is common in Bayesian-based GSI analyses (Pella & Masuda 2001). Models fit by ADMB-ML or ADMB-PMLE were considered to have converged on a solution when the maximum gradient of the parameters with respect to the objective function was <0.1E–4.

Both SPAM and ONCOR have several options available for analysis. For SPAM, we used iteratively reweighted least squares and without resampling of the source populations or mixtures. The Rannala & Mountain (1997) approach for calculating allele relative frequencies of the source populations was used. Termination criteria when fitting models in SPAM (see ADFG 2000) were the following: estimate tolerance = 0.1E–20; likelihood tolerance = 0.1E–20; and guaranteed percentage achievement of the maximal likelihood = 90%. For ONCOR, resampling was not conducted on the source populations or mixtures, and default termination criterion was used (see Kalinowski, Manlove & Taper 2007). ONCOR uses the Rannala & Mountain (1997) approach to calculate source population allele relative frequencies (Kalinowski, Manlove & Taper 2007).

SIMULATIONS AND EVALUATED FACTORS

Because SPAM and ONCOR have graphical user interfaces, we used AutoHotkey (www.autohotkey.com), which is a graphical user interface scripting program, to record executable files of the keystrokes needed to run these programs. A ‘driver’ program (i) looped over the factor-level combinations and repeatedly called the simulation program to produce a data set, (ii) called each of the estimation routines to analyse the simulated data set, and (iii) compiled and appended results from the estimation routines into a single output file. The ‘driver’ program was also written using AD Model Builder but was separate from both the simulation and estimation programs.

The following factors were evaluated in this study: number of source populations (three levels: 5, 20 and 40 populations), source population divergence ($\theta$) (three levels: 0.01, 0.06 and 0.15), number of evaluated loci (five levels: 5, 10, 15, 50 and 100 loci), source population sample size (three levels: 50, 100 and 200 fish) and mixture sample size (four levels: 100, 250, 500 and 1,000 fish). We used a full factorial design so that all $3 \times 3 \times 5 \times 3 \times 4 = 540$ combinations of factors levels were explored with 500 simulations conducted for each combination. The expected contribution patterns were similar for each of the assessed source population levels and reflected situations in which a few populations contributed most of the individuals to a mixture. The assumed contribution pattern was such that 20% of the populations contributed c. 60, 24, 10, and 4% of the individuals to the mixture with the maximum and minimum contributions for individual populations depending on number of source populations (Fig. 1). For simulations with 50 or 100 loci, the number of alleles per locus was set at two to mimic SNP markers. For simulations with 15 or fewer loci, the number of alleles was randomly selected for each locus and iteration and could range from 5 to 25 alleles.

There is considerable variability in the conditions under which GSI or other analyses involving population structure are performed. Number of source populations have ranged from fewer than 5 (Bott et al. 2009) to more than 200 (Beacham et al. 2005, 2006), although when the number of source populations is large, they are often grouped into fewer regional reporting units (Ackerman, Habicht & Seeb 2011; Clemento et al. 2014; Larson et al. 2014). Likewise, genetic divergences, number of evaluated loci, number of alleles per locus, source population sample size and mixture sample size can vary widely [genetic divergence: $<0.01$ (Small et al. 2005; Larson et al. 2014) to $>0.20$ (Clemento et al. 2014); number of loci: $<10$ (Smith et al. 2005; Habicht, Seeb & Seeb 2007) to $>200$ (Moen et al. 2008); number of alleles per locus: 2 (Bradbury et al. 2011; Clemento et al. 2014) to $>40$ (Beacham et al. 2005, 2006; Habicht, Seeb & Seeb 2007); source population sample size: $<50$ (Smith et al. 2005; Bradbury et al. 2011; Clemento et al. 2014) to $>300$ (Hasselman, Bradford & Bentzen 2010; Beacham et al. 2012); and mixture sample size: 100 or fewer (Bott et al. 2009; Chapman, Pinhal & Shivji 2009) to $>2000$ (Beacham et al. 2012; Clemento et al. 2014)]. We attempted to capture this level of variability in our factor-level selections, while allowing simulations to be completed in a timely manner. Our choice of factor levels was also influenced by our belief that differences in estimation routines would likely be greater at
conditions considered less informative for GSI analyses (e.g. low genetic divergences, fewer evaluated loci, small source population and mixture sample sizes).

**PERFORMANCE MEASURES**

Aitchison distance (AD; Aitchison 1986) was used to measure agreement between actual and estimated source population contributions. AD was calculated as

\[
AD(p, \hat{p}) = \sqrt{\sum_{i=1}^{n} \left( \log_e p_i / \prod_{k=1}^{n} p_k - \log_e \hat{p}_i / \prod_{k=1}^{n} \hat{p}_k \right)^2},
\]

where \(p_i\) and \(\hat{p}_i\) are actual and estimated contributions for the \(i\)th source population, respectively. Because rounded values of zero for the estimated contributions would cause an error in the AD calculations, zero estimates were replaced with a small constant (0.0001) and the nonzero estimates were adjusted multiplicative to maintain the unit-sum constraint of the contributions (Martín-Fernández, Palarea-Albaladejo & Olea 2011).

Multifactor ANOVA models were fit to the ADs from the simulations for assessing the relative importance of the investigated factors. Analyses were conducted separately for each source population level. We used \(\eta^2\) values as estimates of the proportion of explained variability in the response variable accounted for by main effects and interactions (Corell et al. 2012). The median of the ADs from the simulations conducted for a particular combination of factor levels was used as a measure of estimation routine accuracy, while the interquartile range (IQR) of the ADs was used to measure consistency. To facilitate comparison, estimation routines were ranked with respect to each other for both AD medians and IQRs for each investigated scenario, with lower rankings indicative of better performance.

**Results**

Based on effect sizes (\(\eta^2\) values) from the multi-factor ANOVA models, mixture sample size and source population genetic divergence accounted for most of the explained variability in the results (see Tables S1 and S2, Supporting information for \(\eta^2\) values) when all four estimation routines were included in the model. These main factor effects together accounted for between 66 and 78% in explained variability in the ADs for the different source population levels. For simulations involving five or 20 source populations, mixture sample size accounted for most of the explained variability in ADs, while for simulations involving 40 populations, source population genetic divergence accounted for most of the explained variability. Next to mixture sample size and source population genetic divergence, estimation routine accounted for the next greatest amount of explained variability, with the proportion of explained variability ranging from 8 to 13%. Interactions between estimation routine and other main factor effects were generally small, in most cases accounting for <2% of the explained variability in the ADs, suggesting that performance of the estimation routines was consistent among levels of the other factors. The exceptions to this were for simulations involving five source populations where the interaction between estimation routine and mixture sample size and estimation routine and source population divergence accounted for c. 4% of the explained variability in ADs, suggesting there was some inconsistency with respect to how estimation routines performed for different levels of these factors. Number of evaluated loci and source population sample size had the smallest effect sizes of the main factor effects. Number of evaluated loci accounted for between 5 and 7% of the explained variability in the ADs depending on the number of source populations, whereas source population sample size generally accounted for <0.1% of the explained variability.

To assess consistency of the conditional maximum likelihood estimation routines, we repeated the ANOVA effect size calculations but excluded the ADMB-PMLE results. Mixture sample size and source population genetic divergence continued to account for the greatest amount of explained variability in the simulation ADs (see Table S2). Proportion
of explained variability by these main effects together ranged from 87 to 88%. The exclusion of ADMB-PMLE resulted in the proportion of explained variability in simulation ADs due to estimation routines declining to <1% for all source population levels.

For simulations involving five source populations, accuracy of the conditional maximum likelihood estimation routines was similar (Fig. 2) as the ANOVA results suggested (median and interquartile ranges of the ADs and rankings of the estimation routines with respect to each other are provided in Table S3, Supporting information). Across all investigated factor levels, median ADs between SPAM, ONCOR and ADMB-MLE differed on average by <1.5%. Although results between the conditional maximum likelihood estimation routines were similar, in most investigated scenarios, SPAM had the smallest median ADs, followed by ONCOR and ADMB-MLE. Across all evaluated scenarios, the average rank of median ADs was 1.73, 2.22 and 2.32 for SPAM, ONCOR and ADMB-MLE, respectively. The accuracy of ADMB-PMLE was noticeably poorer than the other routines for scenarios when mixture sample size and source population genetic divergence were low (Fig. 2), although this was least evident at the highest divergence level (θ = 0.15). Accuracy of ADMB-PMLE approached those of the conditional maximum likelihood estimates as mixture sample sizes increased (Fig. 2). These variations in the relative performance of ADMB-PMLE likely explain why interactions between estimation method and either genetic divergence or mixture sample size explained a non-trivial amount of the variation for simulations involving five source populations. While higher order interactions did not explain a substantial amount of variation overall, there were some noteworthy results. Although the accuracy of ADMB-PMLE generally approached that of the other methods for large mixture sample sizes, there were variations as to when this occurred. In particular, for the lowest genetic divergence (θ = 0.01) and largest source population sample size, ADMB-PMLE generally required larger mixture sample sizes for it to approach the level of accuracy of the other routines. As well, for simulations involving 50 loci, even at the largest mixture sample size, there were noticeable discrepancies in the ADMB-PMLE estimates when compared to the other routines (Fig. 2).

In terms of consistency for simulations involving five source populations, results were slightly more variable (Fig. 3). The consistency of SPAM, ONCOR and ADMB-MLE was fairly similar regardless of mixture sample size in many of the examined scenarios (Fig. 3). The primary exceptions to this were at lower genetic divergences (θ = 0.01 or 0.06). Depending on the number of evaluated loci and source population sample sizes, the conditional maximum likelihood estimation routines approached similar levels of consistency at mixture sample sizes ranging from 100 to 500 fish. For consistency, SPAM most often had the smallest IQR of the ADs, followed by ADMB-MLE, ONCOR and ADMB-PMLE. Across all evaluated scenarios, the average rank in terms of the IQR of the ADs was 1.63, 2.23, 2.35 and 3.79 for SPAM, ADMB-MLE, ONCOR and ADMB-PMLE, respectively. On average, the IQRs of the ADs for SPAM, ONCOR and ADMB-MLE differed by ca. 8%. In case of ADMB-PMLE, consistency of the estimates approached those of the conditional maximum likelihood estimation routines at mixture sample sizes ranging from 250 to 1000 depending on the combination of factor levels, following patterns similar to those seen for the accuracy results (Fig. 3).

Fig. 2. Median Aitchison distance for simulations involving five source populations by estimation routine. The different panels correspond to different combinations of genetic divergence between source populations (Div.) and source population sample size (Pop. Size). The x-axis identifies number of evaluated loci and mixture sample size.
The results for the simulations involving 20 or 40 source population levels were similar with respect to accuracy. For all estimation routines, accuracy clearly improved as mixture sample size and source population genetic divergence increased (Figs 4 and 5). Although the conditional maximum likelihood estimates exhibited overall similar levels of accuracy across most of the investigated scenarios, ADMB-PMLE often exhibited poorer accuracy compared to the other estimation methods for many of the factor-level combinations (Figs 4 and 5). In some instances [e.g. source populations = 40; source population sample size = 50; genetic divergence (0) = 0.01], the accuracy of ADMB-PMLE was closer to that of the other routines at lower mixture sample sizes than at larger sample sizes; in other cases (e.g. source populations = 20; source population sample size = 50; genetic divergence (0) = 0.06), the amount of bias in ADMB-PMLE was fairly consistent across mixture sample sizes. Similar to simulations involving five source populations, although the conditional maximum likelihood estimation routines exhibited overall similar levels of accuracy across most of the investigated scenarios, ADMB-PMLE often exhibited poorer accuracy compared to the other estimation methods for many of the factor-level combinations (Figs 4 and 5). In some instances, the accuracy of ADMB-PMLE was closer to that of the other routines at lower mixture sample sizes than at larger sample sizes; in other cases, the accuracy of ADMB-PMLE was fairly consistent across mixture sample sizes.

In terms of consistency of the estimation routines for scenarios involving 20 or 40 source populations, the results for ADMB-PMLE were highly variable. In some scenarios, ADMB-PMLE had the lowest average ranking for simulations involving 20 (average rank = 1.74) and 40 (average rank = 1.82) source populations. However, closer examination of the ADMB-PMLE estimates in scenarios where it had the best consistency indicated this largely was a consequence of it consistently underestimating contributions for some of the source populations meaning it did not necessarily reflect better performance. The conditional maximum likelihood estimation routines again generally exhibited similar levels of performance overall (Figs 6 and 7). There were some instances where one of the routines clearly exhibited better consistency than the others, but differences from the other routines were generally small compared to the differences with the ADMB-PMLE routine. For simulations involving 20 source populations, ONCOR generally had the best consistency followed by SPAM and ADMB-MLE. The average rankings of ONCOR, SPAM and ADMB-MLE with respect to the IQR of the ADs were 1.24, 2.29 and 2.71. For 40 source populations, the average rankings of SPAM, ONCOR and ADMB-MLE were 1, 2.49, 2.69 and 2.88, respectively. For simulations involving 40 source populations, ONCOR again generally had the best consistency followed by AMB-MLE and SPAM. The average rankings of ONCOR, ADMB-MLE and SPAM with respect to the IQR of the ADs were 2.49, 2.69 and 3.01, respectively.

**Discussion**

Given that GSI analyses provide important information for tasks such as diagnosing illegal harvest or overharvest of vulnerable source populations, it is important that contribution estimates be as accurate as possible (Anderson, Waples & Kalinowski 2008; Ensing et al. 2013). In general, we found that conditional maximum likelihood estimation routines (SPAM, ONCOR and ADMB-MLE) performed similarly across a...
range of evaluated scenarios. This was evident based on both plots of the medians and IQRs of the ADs but also by effect sizes from ANOVA models fit to the simulation results, which indicated that only a small fraction of explained variability could be explained by estimation routine. Despite the overall level of similarity between the conditional maximum likelihood estimation routines, in most examined scenarios, the contribution estimates from SPAM were marginally more accurate than those from ONCOR or ADMB-MLE, with ONCOR generally exhibiting better performance than the ADMB-MLE routine. There were some scenarios for which ONCOR and ADMB-MLE exhibited the best performance in terms of accuracy, but SPAM had the best accuracy for more than 75% of the examined scenarios. Conversely, ADMB-

Fig. 4. Key same as in Fig. 2 except for 20 source populations.

Fig. 5. Key same as in Fig. 2 except for 40 source populations.
PMLE generally was the least accurate, although at least for some source population levels as mixture sample size and source population genetic divergence increased the ADMB-PMLE estimates approached those of the other routines.

The use of SNPs in genetic analyses of population structure and diversity has become common and is anticipated to increase as biologists and managers become more aware of their beneficial features (Ogden 2008). In the Pacific Northwest of North America, SNP panels are routinely being used to examine population structure and diversity in salmonid species, including sockeye salmon *Oncorhynchus nerka* (Ackerman, Habicht & Seeb 2011; Creelman et al. 2011), chinook salmon *O. tsawytscha* (Smith et al. 2005; Larson et al. 2014), steelhead *O. mykiss* (Limborg et al. 2012) and chum salmon *O. keta* (Garvin et al. 2013). In terms of incorporating SNPs in GSI analyses, our results from scenarios that were intended to mimic these markers suggest that the estimation routines evaluated in our research exhibited similar levels of performance with respect to each other when compared to scenarios that mimicked other marker types. In particular, the conditional maximum likelihood estimation routines had similar levels of performance although SPAM generally exhibited slightly better performance than either ONCOR or ADMB-MLE. Initially, we hypothesized that features of SNP markers could affect relative performance of different GSI estimation routines. This hypothesis was not supported for the conditional maximum likelihood implementations, and thus, there seems no need to switch between programs depending on marker type incorporated in GSI analyses.

Although accuracy of the ADMB-PMLE routine in some instances approached those of the conditional maximum likelihood estimation routines, in many of the examined scenarios, there were noticeable biases in source population contribution estimates even when genetic divergences and mixture sample sizes were large. This method seemed to perform especially poorly under conditions intended to mimic use of SNP markers (50 or 100 loci and 2 alleles). It is possible that if we had used different prior structures or numerical search algorithms (e.g. expectation-maximization algorithm) or incorporated a regularization coefficient that controlled the trade-off between the penalty and model fit, then performance of the ADMB-PMLE routine might have been better (Ciuperca, Ridolfi & Idier 2003). Consequently, we caution that our results should not be interpreted as a general condemnation against penalized maximum likelihood estimation with respect to GSI analyses. However, it is clear that relatively poor performance results when simply adding a penalty based on a prior probability distribution commonly incorporated in Bayesian-based GSI analyses (Pella & Masuda 2001) and using standard numerical methods implemented in AD Model Builder. We recommend that alternative approaches to using penalized likelihood for GSI analyses be evaluated with simulations prior to adoption.

As previously indicated, penalized maximum likelihood estimation has a Bayesian interpretation with the point estimates from the fitted model being equal to the mode of the posterior probability distributions for the parameters. Corander, Marttinen & Måntyniemi (2006) noted that one potential problem with Bayesian GSI analyses is that when there are few individuals in the mixture from some source populations, the posterior probability distributions for these contributions can have a high degree of uncertainty. This may explain the poor performance of the ADMB-PMLE routine relative to the other programs. In this research, some populations would be expected to contribute few if any fish to a mixture sample (especially for

high numbers of populations and lower mixture sample sizes).
It is possible that if we had included a fully Bayesian-based
estimation routine, accuracies of the source population contribu-
tions, based on say the mean of the posterior distributions, could have improved relative to the conditional maximum likelihood estimation routines. While it would be advantageous to include Bayesian-based methods in comparisons with GSI estimation routines, the main challenge is in determining how to conduct these comparisons efficiently. One way to facilitate such comparisons would be for estimation programs to have options for command-line interfaces that could be run on Unix operating systems, which would allow comparisons to be conducted on high-performance computer clusters.

Our finding that estimation routines implemented in specialized software such as SPAM and ONCOR typically performed better than conditional maximum likelihood estimation routines programmed in AD Model Builder suggests that those interested in performing GSI analyses need not take the time and effort to program their own estimation routine. Perhaps it is not surprising that SPAM and ONCOR performed better than our ADMB-MLE routine given that these programs are specialized for GSI analyses and incorporate methods designed specifically for these types of models (Pella, Masuda & Nelson 1996). We nevertheless thought it would be beneficial to try given the speed and generally good performance of AD Model Builder in estimation (Fournier et al. 2012). Conceivably, if we incorporated features such as those identified by Pella, Masuda & Nelson (1996), then our ADMB-MLE routine would closer match the performance of SPAM and ONCOR, although some of those features might require changing AD Model Builder source code (e.g. modification of the numerical search routine), which would be possible since the software is open source. Despite our finding that SPAM and ONCOR generally had better accuracy than our ADMB-MLE routine, we can still envision some situations where it may be desirable for users to program their own GSI estimation routine. In particular, there can be cases where the existing ‘canned’ GSI programs do not have sufficient flexibility to address questions of interest. For example, Okuyama & Bolker (2005) modelled contributions of loggerhead Caretta caretta and green turtle Chelonia mydas rookeries to large feeding aggregations where the contributions were modelled as functions of rookery size and location within major ocean currents, which necessitated programming their own estimation routine in a different software package. Modifying existing programs to account for this complexity might be possible if source code was available, but without this access, users would need to develop their own routines. The results of our research suggest that estimation routines can be developed and expected to have fairly similar levels of estimation accuracy to existing specialized GSI programs under many conditions although some trade-off in accuracy can be expected without substantial efforts in adapting search routines for the specifics of the GSI problem.

Acknowledgements
This study was partially funded by Great Lakes Fishery Trust project 2009.1080. Additional funding was provided by contributing partners of the Michigan State University Quantitative Fisheries Center.

Data accessibility


