Fisheries Research 193 (2017) 217–222

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

# Incorporating non-baseline characters into genetic mixture analyses

Milo D. Adkison*, Keith R. Criddle

*College of Fisheries and Ocean Sciences, University of Alaska Fairbanks, 17101 Pt. Lena Loop Rd., Juneau, AK 99801, USA*

ARTICLE INFO

Handled by Prof. George A. Rose

*Keywords:*
Population mixtures
Mixture analysis
Bayesian statistics
Genetic analysis
Genetic baseline

ABSTRACT

In a mixture of individuals from different populations, population proportions and individual identities are estimated by comparing the characteristics of individuals in the mixture to a (usually) genetic baseline of population-specific characteristics. Using simulated data sets, we examined the performance of a genetic mixture analysis that incorporated data on non-baseline character state frequencies. Population-specific state frequencies of non-baseline characters were well-estimated in many scenarios. We found benefits of incorporating non-baseline characters in mixture analysis; both individual assignments and estimates of population proportions were improved. However, both the sample size and the quality of the baseline data were more important. We did not see any improvement in estimating baseline character state frequencies even when highly informative non-baseline data was used. Our results suggest that non-baseline data might improve mixture analyses, and we note that population-specific estimates of non-baseline character state frequencies are often useful in and of themselves.

## 1. Introduction

The Bayesian mixture analysis estimation methodology developed by Pella and Masuda (2001) uses a baseline of character state frequencies (such as the frequency of a specific allele at a locus) in each population to provide probability distributions for the proportions of each population in a mixture. As a part of its calculation methodology, it also provides the probability that an individual in a mixture belongs to a particular population. One novel aspect of this particular Bayesian approach is that rather than simply making inference about the mixture from baseline data, it acknowledges that the baseline data also comes from a sample that may not be fully representative of the underlying population; it then uses data from the mixture to improve the estimates of the character state frequencies in each population. That is, instead of thinking of this methodology as a way to estimate proportions in a mixture, it can instead be viewed as a way to use mixture data to help estimate population characteristics.

This leads to several hypothetical questions. First, could this approach be used to estimate the frequency in a population of alternative states of characters for which there are no baseline data? For example, salmon populations that migrate to sea and are caught in a mixed-stock fishery might differ in age or length frequencies when they are caught (Larson et al., 2013; Myers et al., 2007). These age and length frequencies at the time and location where the fishery occurs would not be a part of the baseline data, since baseline data are collected from fish of previous generations on the spawning grounds (Guthrie et al., 2015; Seeb et al., 2007). A few recent studies have demonstrated the practicality of estimating the population-specific frequencies of non-baseline character states (e.g., Moran et al., 2014; Tsehaye et al., 2016).

Second, are these non-baseline character states useful for better characterizing the origin of an individual organism in a mixture? Such an improvement would be quite helpful – large samples from a mixture are required for estimating population frequencies, often forcing aggregation of samples from large areas and long periods of time. The resulting coarse spatio-temporal resolution limits our ability to explore questions about fine scale population distribution and migratory patterns. For some management purposes, such as enforcing endangered species protections, determining what population an individual originated from is essential (Nielsen et al., 2012; Ogden and Linacre, 2015).

Either case seems reasonable. For example, if a population is characterized by a smaller than average size, it seems intuitive that noting that an individual in a mixture is small should increase our certainty that it is a member of that population. However, it's also plausible that the information provided by size is "used up" in estimating the population-specific size distributions, resulting in no improvement in estimating the origin of individuals.

Finally, assuming the state frequencies of characters not sampled in the baseline could be estimated, would these characters then be useful for better characterizing the makeup of the population mixture? For example, could one use the age or length of an individual salmon

* Corresponding author.
*E-mail addresses:* mdadkison@alaska.edu (M.D. Adkison), kcriddle@alaska.edu (K.R. Criddle).

http://dx.doi.org/10.1016/j.fishres.2017.04.016
Received 5 December 2016; Received in revised form 17 April 2017; Accepted 24 April 2017
Available online 04 May 2017

caught in a mixed-stock fishery to better ascertain its identity, and thus improve estimates of the proportion of each population in the mixture?

In this study, we use simulated data to examine under which circumstances state frequencies of a non-baseline character can be estimated using data from a mixture, whether using such characters improves estimates of baseline character state frequencies, and when using a non-baseline character in a mixture analysis improves estimates of population proportions and/or increases the accuracy of assignment of individuals to their population of origin.

## 2. Methods

### 2.1. Simulated data

We simulated baseline data for four populations with two independent baseline characters. The first character had four possible states, and frequencies differed among each population. The second character had two states, and pairs of populations had identical frequencies, mimicking a regionally-varying character. We simulated baseline data by randomly generating state frequencies for each character from each population. Each character's baseline sample state frequencies were determined by generating a random draw from a Dirichlet distribution whose parameters were the product of the true frequencies and different sample sizes.

We then simulated a mixture where 70% of the individuals came from one population and 10% each came from the other three populations. Each individual in the mixture had character states drawn randomly from its population's true character state frequencies. In addition to the two characters contained in the baseline, each individual was assigned a state for another independent character for which there was no baseline data. There were four states for this character, and state frequencies differed among the four populations.

### 2.2. Scenarios investigated

We created scenarios that differed in: the number of individuals sampled in each population to create the baseline (20, 100, 500), number of individuals sampled in the mixture (also 20, 100, and 500), the contrast among populations in state frequencies of the two baseline characters (Table 1), and the contrast among populations in state frequencies of the non-baseline character (Table 1). These scenarios are abbreviated in Figures using the sample size followed by two letters, the first of which gives the contrast in the baseline characters and the second that of the non-baseline character. For example, "100LH" indicates that sample sizes (both baseline and mixture) were 100, that baseline characters had low contrast, and that the non-baseline character had high contrast.

**Table 1**
State frequencies for each character at each level of contrast.

| Contrast level | State values |
|---|---|
| low baseline | Character 1: frequency of state i = 0.4 in population i, = 0.2 in other populations<br>Character 2: state 1 = 0.67 in populations 1–2, 0.33 in populations 3–4<br>state 2 = 0.33 in populations 1–2, 0.67 in populations 3–4 |
| high baseline | Character 1: frequency of state i = 0.7 in population i, = 0.1 in other populations<br>Character 2: state 1 = 0.9 in populations 1–2, 0.1 in populations 3–4<br>state 2 = 0.1 in populations 1–2, 0.9 in populations 3–4 |
| low non-baseline | frequency of state i = 0.4 in population i, = 0.2 in other populations |
| high non-baseline | frequency of state i = 0.7 in population i, = 0.1 in other populations |
| perfect non-baseline | frequency of state i = 1.0 in population i, = 0.0 in other populations |

### 2.3. Computation

For each scenario, we simulated 1000 sets of data. We applied a slightly modified version of the Pella-Masuda Bayesian estimation methodology (2001) to each dataset, and estimated both the proportion of each population in the mixture and the frequencies of alternative states of each character in each population. The posterior distributions of the estimates were compared to the true values. At each iteration of the MCMC calculation in the Pella-Masuda methodology, each individual in the mixture is assigned a population identity (see below); after convergence, we tracked the frequency of assignment of the simulated individuals to the correct population. We tracked how well the state frequencies of the non-baseline character were estimated, how well the state frequencies of the baseline characters were estimated, and whether and to what extent using an informative non-baseline character improved estimates of baseline frequencies and assignment of individuals in the mixture to their population of origin.

The Bayesian statistical model of the data and parameters was as follows:

The baseline data $Y = [y_{ijh}]$, where $y_{ijh}$ is the count of state $h$ of character $j$ in the baseline sample of size $n_i$ from population $i$.

$y_{ij.} \sim$ multinomial($n_i$, $q_{ij.}$), where $q_{ijh}$ is the true frequency of state $h$ of character $j$ in population $i$.

$(q_{ij1}, q_{ij2,\ldots}) \sim$ Dirichlet($\beta_{j1}, \beta_{j2,\ldots}$), under the assumption that state frequencies exhibit some degree of similarity among populations (this assumption was not true for our simulated data, but is a plausible assumption in most real-world applications).

Simplifying Pella and Masuda's (2001) approach, we set a weakly informative prior for the $q$'s for character $j$ as a Dirichlet distribution, with the value of its parameters $\beta_{jh}$ equal to the unweighted average of the sampled state frequencies across all populations (i.e., $\Sigma_{h\beta jh} = 1$). For the non-baseline character, the parameter values were set to $1/H$, where $H$ was the number of states for the character.

The mixture data $X = [x_m]$, where $x_m$ is the "genotype", or set of character states of individual $m$ in the mixture.

$$\Pr(x_m \textit{ comes from stock } i) \text{ is proportional to } p_i \times \Pr(x_m \mid \textit{stock } i)$$

$\Pr(x_m \mid \textit{stock } i) = q_{i1m} \times q_{i2m} \times \ldots$ (if continuous characters are involved, the frequency is replaced by the probability density for the observed state value of the character (Bromaghin et al., 2011)).

Following Pella and Masuda (2001), an uninformative prior for the $p$'s was Dirichlet ($1/I$, $1/I$, …), where $I$ is the total number of populations.

Computation of the MCMC sample from the posterior distributions was accomplished with a Gibbs sampler, which involves a sequence of draws from distributions of parameters conditional on the current values of the other parameters. Computation was simplified by using a data augmentation step (Gelman et al., 2014; Pella and Masuda, 2001). At each iteration of the MCMC algorithm, individuals in the mixture were assigned a population of origin by random draw based on the current probabilities an individual with their character states originated from each population. Thus, each iteration of the Gibbs sampler consisted of the following steps:

1. Assign a random population identity to each individual in the mixture sample, where the probability of assignment to population $i$ is proportional to the current value of $p_i \times \Pr(x_m \mid \textit{stock } i)$.
2. Draw random values for the proportion of each population ($p_i$) in the mixture from a Dirichlet distribution where the $i$-th parameter = $1/I$ + the count of all individuals in the mixture assigned to population $i$.
3. Draw random values for the population-specific state frequencies of all characters, baseline and non-baseline, where the frequency of