Incorporating non-baseline characters into genetic mixture analyses

Milo D. Adkison* and Keith R. Criddle

College of Fisheries and Ocean Sciences, University of Alaska Fairbanks

17101 Pt. Lena Loop Rd.

Juneau, AK 99801 USA

(907) 796-5441 fax 796-5447 mdadkison@alaska.edu, kcriddle@alaska.edu

* corresponding author. E-mail address: mdadkison@alaska.edu

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.

Highlights:

- Population-specific differences in non-baseline characters can be estimated from a mixture
- Non-baseline characters only slightly improved estimates of population proportions in a mixture
- Non-baseline characters are more useful in assigning population identities to individuals
- Non-baseline characteristics may be useful in other ways; e.g., age and size is related to mortality
- Individual assignment allows better spatio-temporal resolution than mixture analysis

keywords: population mixtures, mixture analysis, Bayesian statistics, genetic analysis, genetic baseline

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 and others 2013; Myers and others 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 III and others 2015; Seeb and others 2007). A few recent studies have demonstrated the practicality of estimating the population-specific frequencies of non-baseline character states (e.g., Moran and others 2014; Tsehaye and others 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 and others 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 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.

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.

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

$(q_{ij1}, q_{ij2}, \ldots) \sim \text{Dirichlet}(\beta_{j1}, \beta_{j2}, \ldots)$, 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 β_{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.

$\Pr(x_m \textit{ comes from stock } i)$ 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 and others 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 and others 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 $p_i \times \Pr(x_m \mid stock\ i)$.
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 state h of character j in population i is drawn from a Dirichlet with the h-th parameter = $\beta_{jh} + y_{ijh}$ + count of state h in all mixture individuals assigned to population i.

Based on preliminary trials, we found that 1000 iterations of the MCMC algorithm were sufficient to achieve convergence (Gelman's R << 1.1). Accordingly, each MCMC chain was run for 2000 iterations, and inference was based on the last half of the series.

3. Results

The ability to estimate population-specific frequencies of the states of a non-baseline character was affected both by the sample size and by the degree of contrast in the baseline character state frequencies.

151 At sample sizes of 500, the estimates of non-baseline state frequencies were almost identical to the true
152 values (Fig. 1a). This was also true with a smaller sample size of 100, as long as the baseline contrast was
153 high. The width of the 90% credible interval was strongly affected by both the sample size (1st part of
154 scenario abbreviation) and the contrast in the baseline characters (2nd letter of scenario, Fig. 1b).

155 Although no scenario showed any bias in estimating frequencies of a baseline character (Fig. 2a), the
156 width of the 90% credible interval was strongly affected by sample size, and was also improved when the
157 baseline character had higher contrast (Fig. 2b). Including a non-baseline character in the analysis did
158 very little to improve estimates of baseline character state frequencies, irrespective of the amount of
159 contrast among stocks in non-baseline state frequencies. Even when the non-baseline character was fixed
160 at different states in different populations, little to no improvement in bias or precision was observed (Fig.
161 2, scenarios ending in "P")

162 Including non-baseline characters did improve the accuracy of population assignments for individuals in a
163 mixture, but only slightly (Fig. 3). The accuracy of individual assignments depended mainly on the
164 contrast in the baseline characters, and to a smaller extent on sample sizes.

165 Under some circumstances, including a non-baseline character also improved the precision (Fig. 4b) of
166 estimates of population proportions in the mixture. For instance, with a sample size of 500 and low
167 contrast in baseline characters, the average width of the 90% credible interval for the proportion of
168 population 1 in the mixture decreased from 0.30 to 0.18 when a non-baseline character that was fixed at
169 different states in different stocks was included (Fig. 4b, 500LN vs. 500LP). The contrast in the baseline
170 characters and the sample size showed larger effects, however. Sample size had a fairly large effect on
171 bias when the baseline contrast was low (left half of Fig. 4a), while the contrast in baseline characters
172 affected both bias and precision (left vs. right half of Fig. 4a&b).

173

4. Discussion

The stock-specific state frequencies of non-baseline characters can be estimated fairly well from mixture data, given adequate sample sizes and contrast in the baseline characters. Non-baseline characters can provide some improvement in the estimate of population proportions in a mixture or in identifying the population of origin of individuals in a mixture. Our results suggest that analysts performing mixture analysis should consider including data on non-baseline characters. However, the improvements resulting from including non-baseline characters are small relative to the effects of sample size or of a baseline with strong contrast among populations. Assembling a comprehensive and informative baseline and obtaining a representative and adequate sample from both baseline individuals and individuals in the mixture of interest should be a high priority. Using non-baseline characters made no noticeable improvement in estimating state frequencies of baseline characters.

The ability to estimate differences among populations in characteristics not present in the baseline can be quite useful for management purposes. Bromaghin et al. (2011) and Moran et al. (2014), in developing the methodology employed here, examined differences in fecundity and disease prevalence among populations. Studies using less sophisticated methodologies (see list in Moran et al. 2014) have looked at an even wider range of characters. Tsehaye et al. (2016) were able to estimate population-specific (relative) recruitment by incorporating age or length data into mixture analysis, but made some strong assumptions about the underlying population dynamics and life histories of the populations contributing to the mixture.

One immediate practical application would be estimating stock-specific ocean size frequencies, a non-baseline character, of Chinook salmon (*Onchoryhnchus tshawytscha*) taken as bycatch in Bering Sea and Gulf of Alaska groundfish fisheries. A recent study of the effect of this bycatch on weak stocks in western Alaska river systems estimated that at its peak, this bycatch reduced returning Chinook abundance by 7% (Ianelli and Stram 2015), although current impacts are much smaller. However, this analysis may have inadvertently overestimated the reduction in western Alaska stocks. Because of the proximity of these

fisheries to the western Alaska populations, bycaught individuals from western Alaska are likely younger than individuals from other populations that make a significant contribution, such as British Columbia or the Pacific Northwest of the United States (Larson and others 2013; Myers and others 2007); these length differences might be informative enough to improve the estimates of the proportions of these stocks in the mixture.

Even if the estimates of proportions were not improved, estimating the population-specific length distributions could still be valuable. Younger fish experience a higher cumulative natural mortality before returning to spawn, so that ignoring the younger age structure of western Alaska fish would result in an overestimate of the bycatch-induced reduction in adult returns to Western Alaska, and an underestimate for other stocks.

While some non-baseline characters may be temporally stable, others, like a population's length distribution, could vary from year to year. For example, Chinook salmon from Western Alaska stocks would consistently have a larger proportion of small individuals in the Bering Sea than stocks from the Pacific Northwest. However, the size distribution would undoubtedly fluctuate from year to year due to inter-annual differences in cohort size and in growth conditions. Such inter-annual differences could easily be incorporated into the estimation procedure as random effects drawn from a hyperdistribution.

One promising result of our simulations was that the probabilistic assignment of individuals to populations was improved by including non-baseline characters. Mixture analysis has been recommended over individual assignment methodologies when the goal is to estimate the proportions in a mixture (Koljonen and others 2005). However, mixture analysis often requires large sample sizes to assign proportions with reasonable uncertainty (Templin and others 2011). This requirement for large sample sizes is problematic when samples are scarce. For example, in investigating the population origins of Chinook salmon bycatch in Bering Sea groundfish fisheries, obtaining adequate sample sizes for mixtures requires aggregating over large areas and long periods, restricting inference to coarse spatial and temporal resolution (Ianelli and Stram 2015). Individual assignments, even if each individual sampled has some

probability of belonging to each of several stocks (or to an unknown stock) (Manel and others 2005), might allow estimation of stock-specific distributions and migration patterns at a finer resolution (Teel and others 2015).

For clarity and to simplify calculations, the simulations in this study produced data that differed significantly from the types of genetic data typically used in most analyses, where there are many more genetic characters, and these characters have many more possible states. Nonetheless, the essentials of the simulation, where populations differed in baseline and non-baseline characteristics, provides qualitative guidance on what non-baseline characters can and cannot contribute to a mixture analysis. A useful follow-up study would be to take a high-quality real dataset and to artificially create non-baseline characters by excluding their baseline data, and to use subsets of baseline characters to create high- and low-contrast baseline datasets.

Acknowledgments

Jeff Guyon of NOAA and Jim Jasper and Sarah Power of the Alaska Department of Fish and Game provided helpful critiques of drafts of this manuscript as did two anonymous reviewers. This work was supported by the North Pacific Research Board (NPRB publication number 636).

242 **References**

243 Bromaghin, J.F., Evenson, D.F., McLain, T.H., Flannery, B.G., 2011. Using a genetic mixture model to
244 study phenotypic traits: Differential fecundity among Yukon River Chinook salmon. Trans. Am.
245 Fish. Soc. 140 (2), 235-249. http://dx.doi.org/10.1080/00028487.2011.558776

246 Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2014. Bayesian Data
247 Analysis. Chapman & Hall/CRC , Boca Raton, FL.

248 Guthrie III, C.M., Nguyen, H.T., Guyon, J.R., 2015. Genetic stock composition analysis of the Chinook
249 salmon bycatch from the 2013 Bering Sea walleye pollock (*Gadus chalcogrammus*) trawl fishery.
250 NOAA Tech. Memorandum NMFS-AFSC-290. doi:10.7289/V5W093V1

251 Ianelli, J.N.; Stram, D.L., 2015. Estimating impacts of the pollock fishery bycatch on western Alaska
252 Chinook salmon. ICES J. Mar. Sci. 72 (4), 1159-1172. https://doi.org/10.1093/icesjms/fsu173

253 Koljonen, M.L., Pella, J.J., Masuda, M., 2005. Classical individual assignments versus mixture modeling
254 to estimate stock proportions in Atlantic salmon (*Salmo salar*) catches from DNA microsatellite
255 data. Can. J. Fish. Aquat. Sci. 62 (9), 2143-2158. doi: 10.1139/f05-128

256 Larson, W.A., Utter, F.M., Myers, K.W., Templin, W.D., Seeb, J.E., Guthrie, C.M., Bugaev, A.V., Seeb,
257 L.W., 2013. Single-nucleotide polymorphisms reveal distribution and migration of Chinook
258 salmon (*Oncorhynchus tshawytscha*) in the Bering Sea and North Pacific Ocean. Can. J. Fish.
259 Aquat. Sci. 70 (1), 128-141. doi 10.1139/cjfas-2012-0233

260 Manel, S., Gaggiotti, O.E., Waples, R.S., 2005. Assignment methods: Matching biological questions with
261 appropriate techniques. Trends in Ecology and Evolution 20 (3), 136-142.
262 http://doi.org/10.1016/j.tree.2004.12.004

263 Moran, P., Bromaghin, J.F., Masuda, M., 2014. Use of genetic data to infer population-specific ecological
264 and phenotypic traits from mixed aggregations. PLoS One. 9 (6), e98470.
265 https://doi.org/10.1371/journal.pone.0098470

266 Myers, K.W., Klovach, N.V., Gritsenko, O.F., Urawa, S., Royer, T.C., 2007. Stock-specific distributions
267 of Asian and North American salmon in the open ocean, interannual changes, and oceanographic
268 conditions. N. Pac. Anadr. Fish Comm. Bull. 4, 159-177.

269 Nielsen, E.E., Cariani, A., Aoidh, E.M., Maes, G.E., Milano, I., Ogden, R., Taylor, M., Hemmer-Hansen,
270 J., Babbucci, M., Bargelloni, L., Bekkevold, D., Diopere, E., Grenfell, L., Helyar, S., Limborg,
271 M.T., Martinsohn, J.T., McEwing, R., Panitz, F., Patarnello, T., Tinti, F., Van Houdt, J.K.J.,
272 Volckaert, F.A.M., Waples, R.S., Albin, J.E.J., Vieites Baptista, J.M., Barmintsev, V., Bautista,
273 J.M., Bendixen, C., Bergé, J.P., Blohm, D., Cardazzo, B., Diez, A., Espiñeira, M., Geffen, A.J.,
274 Gonzalez, E., González-Lavín, N., Guarniero, I., Jeráme, M., Kochzius, M., Krey, G., Mouchel,
275 O., Negrisolo, E., Piccinetti, C., Puyet, A., Rastorguev, S., Smith, J.P., Trentini, M., Verrez-
276 Bagnis, V., Volkov, A., Zanzi, A., Carvalho, G.R., 2012. Gene-associated markers provide tools
277 for tackling illegal fishing and false eco-certification. Nat. Commun. 3, 851.
278 doi:10.1038/ncomms1845

279 Ogden, R., Linacre, A., 2015. Wildlife forensic science: A review of genetic geographic origin
280 assignment. Forensic Sci. International: Genetics. 18, 152-159.
281 http://doi.org/10.1016/j.fsigen.2015.02.008

282 Pella, J., Masuda, M., 2001. Bayesian methods for analysis of stock mixtures from genetic characters.
283 Fish. Bull. 99 (1), 151-167.

284 Seeb, L.W., Antonovich, A., Banks, A.A., Beacham, T.D., Bellinger, A.R., Blankenship, S.M., Campbell,
285 A.R., Decovich, N.A., Garza, J.C., Guthrie, C.M., Lundrigan, T.A., Moran, P., Narum, S.R.,
286 Stephenson, J.J., Supernault, K.J., Teel, D.J., Templin, W.D., Wenburg, J.K., Young, S.E., Smith,
287 C.T., 2007. Development of a standardized DNA database for Chinook salmon. Fisheries. 32
288 (11), 540-552. http://dx.doi.org/10.1577/1548-8446(2007)32[540:DOASDD]2.0.CO;2

289 Teel, D.J., Burke, B.J., Kuligowski, D.R., Morgan, C.A., Van Doornik, D.M., 2015. Genetic
290 identification of chinook salmon: Stock-specific distributions of juveniles along the Washington
291 and Oregon coasts. Mar. Coast. Fish. 7 (1), 274-300.
292 http://dx.doi.org/10.1080/19425120.2015.1045961

293 Templin, W.D., Seeb, J.E., Jasper, J.R., Barclay, A.W., Seeb, L.W., 2011. Genetic differentiation of
294 Alaska Chinook salmon: The missing link for migratory studies. Molecular Ecology Resources.
295 11 (s1), 226-246. 10.1111/j.1755-0998.2010.02968.x

296 Tsehaye, I., Brenden, T.O., Bence, J.R., Liu, W., Scribner, K.T., Kanefsky, J., Bott, K., Elliott, R.F.,
297 2016. Combining genetics with age/length data to estimate temporal changes in year-class
298 strength of source populations contributing to mixtures. Fish. Res. 173 (3), 282-293.
299 http://doi.org/10.1016/j.fishres.2015.09.004

300

301

302 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
	Character 2: state 1 = 0.67 in populations 1-2, 0.33 in populations 3-4
	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
	Character 2: state 1 = 0.9 in populations 1-2, 0.1 in populations 3-4
	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

303

304

305

306

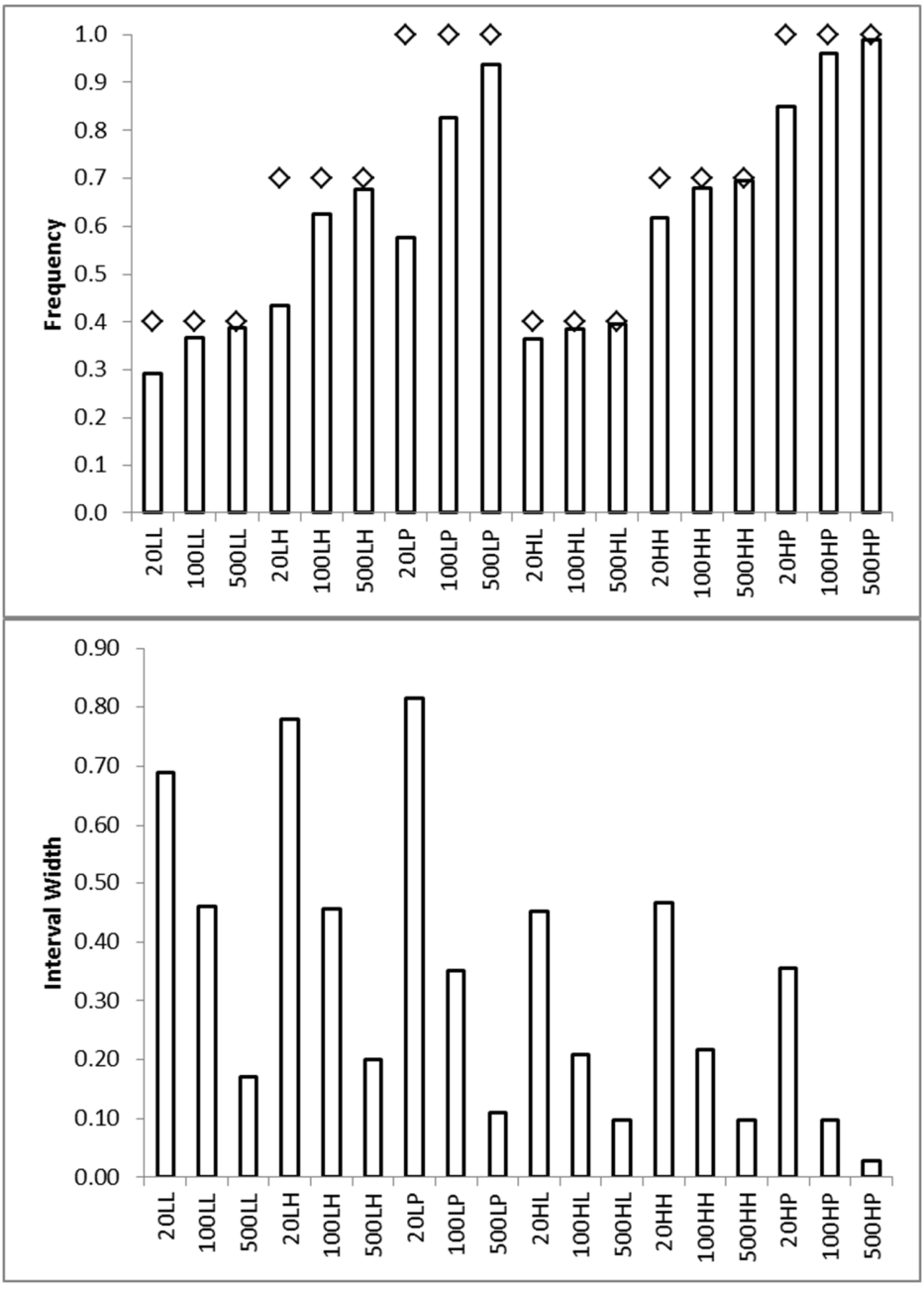
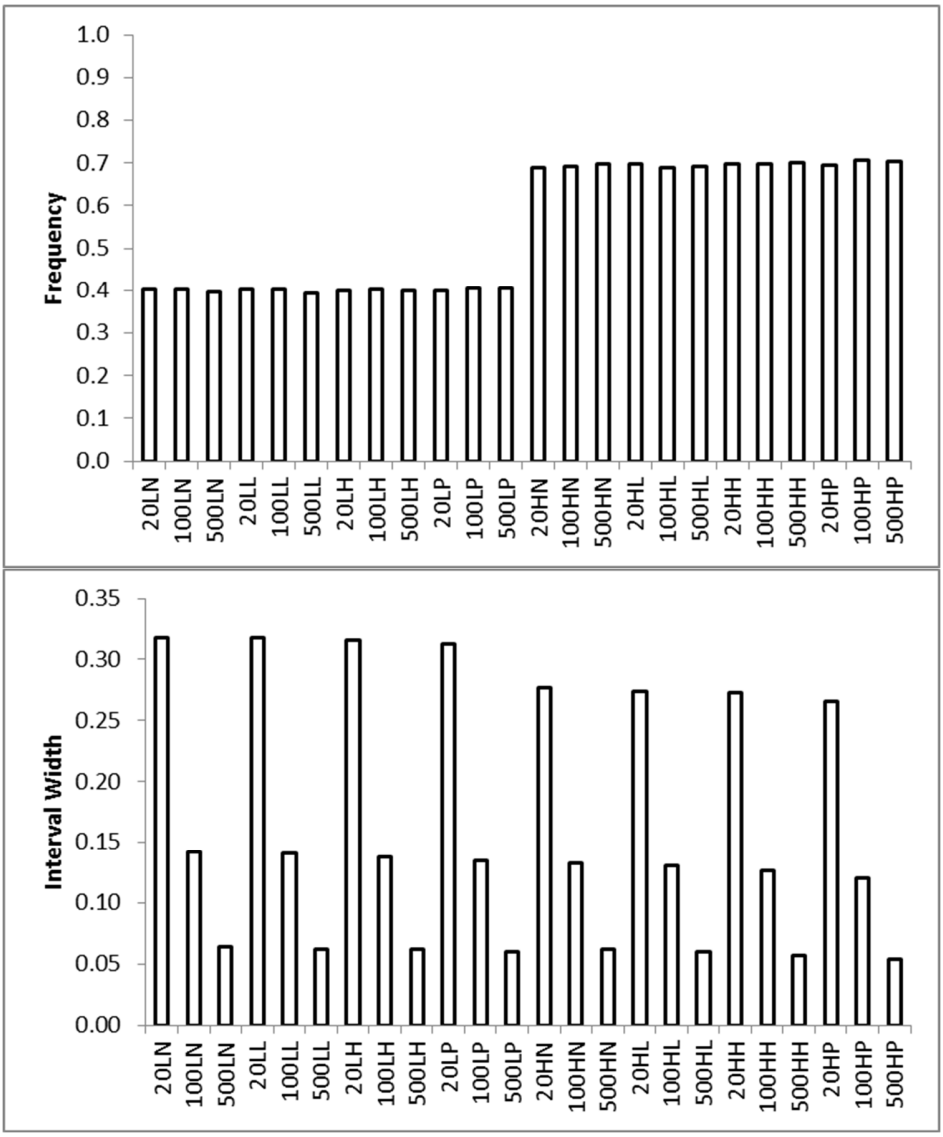


Figure 1. In top figure the bars show the mean estimated frequency of state #1 in the non-baseline character in population #1 from 1000 simulation trials. Diamonds show the true frequency, which was 0.4, 0.7, or 1.0 depending on whether the non-baseline contrast was "L", "H", or "P" (last letter of scenario label). The first two parts of the scenario label on the x-axis indicate sample size (20, 100, 500) and baseline contrast ("L" = low, "H" = high; see Table 1). The bottom figure shows the average width of the 90% credible intervals.

316



317

318 Figure 2. Top figure shows the average estimated frequency of state #1 in the baseline character #1 in
319 population #1 from 1000 simulation trials. The true frequency was 0.4 or 0.7, depending on whether the
320 baseline contrast was "L" or "H" (first letter in scenario label). The first part of the scenario label
321 indicates sample size (20, 100, 500) and the last letter the non-baseline contrast ("N" = no character,
322 "L" = low, "H" = high, "P" = perfect; see Table 1). The bottom figure shows the average width of the
323 90% credible intervals.

324

325

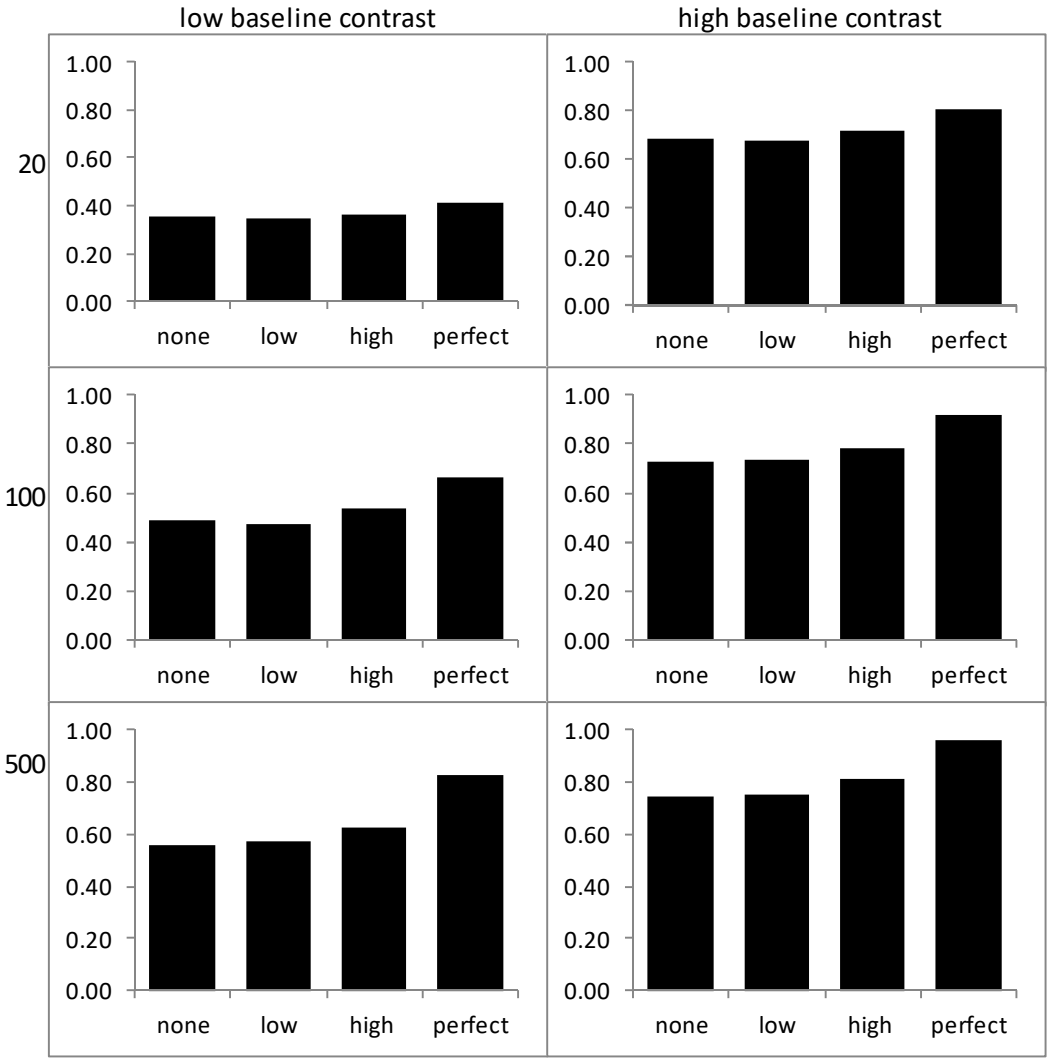


Figure 3. Proportion of individuals in simulated mixtures assigned to the correct population. Left graphs are low baseline contrast, right are high. From top to bottom, graphs are for sample sizes of 20, 100, and 500, respectively. From left to right, columns are with no non-baseline character, then low, high, and perfect contrast in the non-baseline character.

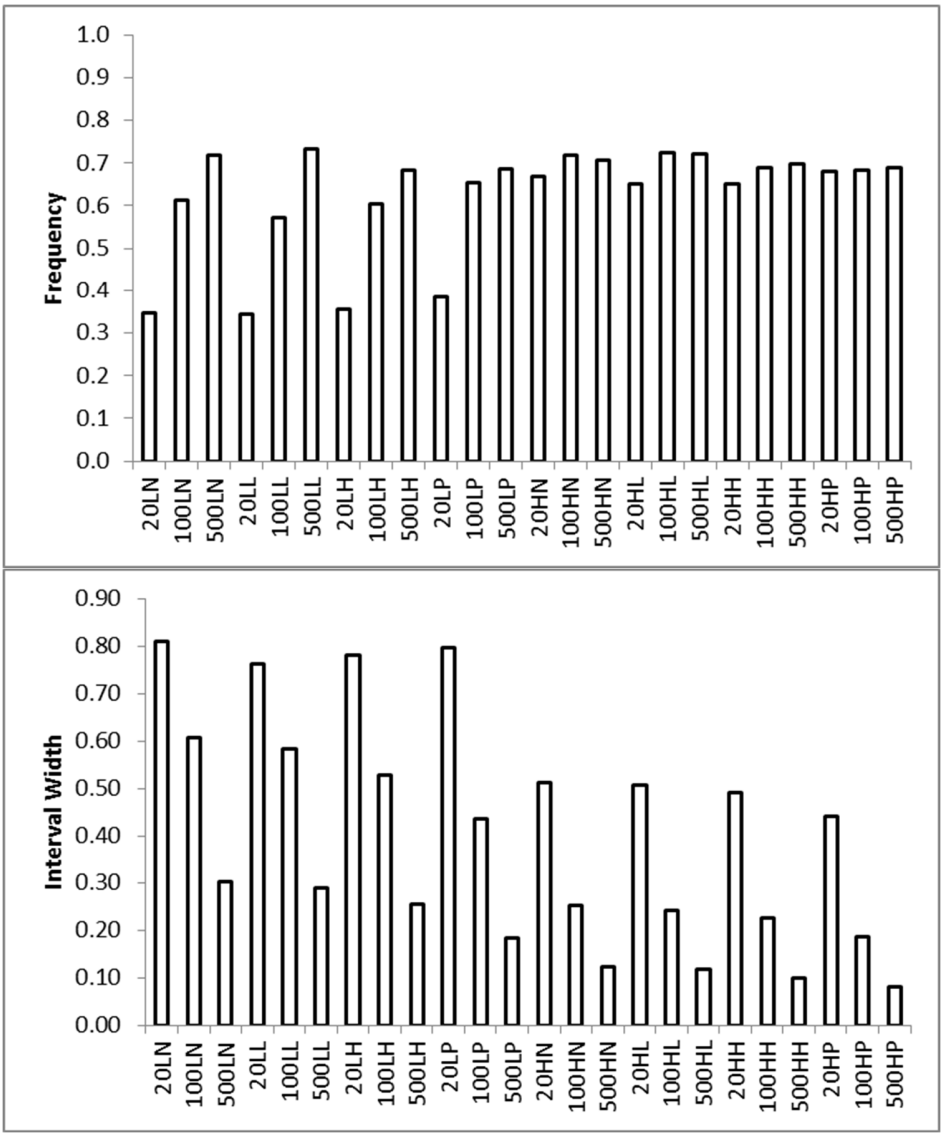


Figure 4. Top graph shows the average estimated frequency of stock #1 in simulated mixtures; the true frequency was 0.7. The first part of the scenario label indicates sample size (20, 100, 500), the first letter the baseline contrast ("L" = low, "H" = high; see Table 1), and the last letter the non-baseline contrast ("N" = no character, "L" = low, "H" = high, "P" = perfect; see Table 1). The bottom figure shows the average width of the 90% credible intervals.