NOAA Technical Memorandum NMFS-AFSC-283

A Bayesian Cross-validation Approach to Evaluate Genetic Baselines and Forecast the Necessary Number of Informative Single Nucleotide Polymorphisms

by
M. R. Garvin, M. M. Masuda, J. J. Pella, P. D. Barry,
S. A. Fuller, R. J. Riley, R. L. Wilmot, V. Brykov, and A. J. Gharrett

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Alaska Fisheries Science Center

November 2014

NOAA Technical Memorandum NMFS

The National Marine Fisheries Service's Alaska Fisheries Science Center uses the NOAA Technical Memorandum series to issue informal scientific and technical publications when complete formal review and editorial processing are not appropriate or feasible. Documents within this series reflect sound professional work and may be referenced in the formal scientific and technical literature.

The NMFS-AFSC Technical Memorandum series of the Alaska Fisheries Science Center continues the NMFS-F/NWC series established in 1970 by the Northwest Fisheries Center. The NMFS-NWFSC series is currently used by the Northwest Fisheries Science Center.

This document should be cited as follows:

> Garvin, M. R., M. M. Masuda, J. J. Pella, P. D. Barry, S. A. Fuller, R. J. Riley, R. L. Wilmot, V. Brykov, and A. J. Gharrett. 2014. A Bayesian cross-validation approach to evaluate genetic baselines and forecast the necessary number of informative single nucleotide polymorphisms. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-AFSC-283, 59 p.
>
> Document available: http://www.afsc.noaa.gov/Publications/AFSC-TM/NOAA-TM-AFSC-283.pdf

NOAA Technical Memorandum NMFS-AFSC-283

A Bayesian Cross-validation Approach to Evaluate Genetic Baselines and Forecast the Necessary Number of Informative Single Nucleotide Polymorphisms

by
M. R. Garvin[1], M. M. Masuda[2], J. J. Pella[2], P. D. Barry[1],
S. A. Fuller[1], R. J. Riley[1], R. L. Wilmot[2], V. Brykov[3], and A. J. Gharrett[1]

[1]Fisheries Division
School of Fisheries and Ocean Sciences
University of Alaska Fairbanks
17101 Point Lena Loop Road
Juneau, AK 99801

[2]Auke Bay Laboratories
Alaska Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
17109 Point Lena Loop Road
Juneau, AK 99801

[3]A.V. Zhirmunsky Institute of Marine Biology
Russian Academy of Science
Far Eastern Federal University
Department of Cell Biology and Genetics
Vladivostok, 690041 Russia

U.S. DEPARTMENT OF COMMERCE
Penny. S. Pritzker, Secretary
National Oceanic and Atmospheric Administration
Kathryn D. Sullivan, Under Secretary and Administrator
National Marine Fisheries Service
Eileen Sobeck, Assistant Administrator for Fisheries

November 2014

This document is available to the public through:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161

www.ntis.gov

ABSTRACT

The determination of the origin of individuals from a mixture composed of multiple populations is becoming a routine tool for the management and study of an increasing number of taxa. It is accomplished by applying statistical methods and a reference genetic baseline whose accuracy and precision must be evaluated to determine its utility. Earlier evaluation methods that used simulated mixtures from a baseline with standard maximum likelihood (ML) methods for mixed-stock analysis (MSA) provided optimistic evaluations of baselines. More recent methods address the optimism but are based solely on ML methods and either do not accommodate potentially informative haploid data or require larger datasets than are available or possible. We used data from a developing baseline for chum salmon (*Oncorhynchus keta*) that includes single nucleotide polymorphisms (SNPs) and microsatellites to produce a method that we call 'leave-ten-percent-out cross-validation' (LTO). This method avoids optimism in baseline evaluation, uses only observed multi-locus genotypes, accepts haploid and diploid data, applies Bayesian methods of MSA, and is less dependent on large baseline sample sizes. In order to further guide the development of genetic baselines, we also simulated increasing numbers of SNP loci and used LTO and logistic regression to estimate the number of informative SNP loci that would be necessary to achieve a specified rate of correct individual assignment.

CONTENTS

INTRODUCTION

Mixed-stock analysis (MSA) uses multi-locus genotype data to estimate the composition of a mixture or to assign its individuals to their respective sources (Pella and Milner 1987, Manel et al. 2005). These types of analyses have been used for the study, management, and conservation of many species (e.g., Wasser et al. 2004, Bowen et al. 2007, Negrini et al. 2008, Griffiths et al. 2010). A reference genetic baseline that includes samples from all of the distinct stocks that may contribute to the mixture is required for MSA. The term "stock" typically refers to a group of individuals defined for conservation or resource management purposes and is not necessarily a breeding population. Regional samples are obtained that represent the likely true geographic range and genetic divergence of the focal species.

Two major goals are to determine the baseline's ability for stock identification and what improvement is needed, if any. Baseline performance is usually evaluated with the baseline samples themselves under the assumption that an analysis of samples from a mixture obtained independently will be similar. The accuracy and precision of the genetic baselines are repeatedly evaluated over time because 1) new loci are often added to a developing baseline to improve performance; 2) new stocks are often added to the baseline to ensure that it is a reliable representation of the complete list of stocks; or 3) only a subset of the baseline is used for MSA, which may be desirable to enhance performance and conserve time and resources.

A widely used method to evaluate genetic baselines is to repeatedly simulate samples of multi-locus genotypes for hypothetical mixtures of a specified stock composition; the compositions of those mixtures are estimated by a conditional maximum likelihood method (ML) with either the actual or simulated baseline samples. The simulations and estimation are accomplished with software programs such as SPAM (Debevec et al. 2000) and GMA

(Kalinowski 2003), both of which overstate the accuracy and precision of baselines (Anderson et al. 2008) because 1) sampling error of baseline allele frequencies inflates the actual genetic divergence among stocks (error-enhanced divergence) and 2) mixtures simulated from the actual baseline samples often have individuals with genotypes identical to those in the baseline (bi-presence). These properties make the correct assignment of individuals to their source stocks optimistic.

The program ONCOR (Anderson et al. 2008) performs baseline evaluation by simulating hypothetical mixtures of genotypes (like SPAM and GMA) but uses a modified ML method for composition estimation that has a leave-one-out (LOO) rule to negate the optimism from bi-presence. However, ONCOR does not accept potentially informative haploid data, whose markers often display greater divergence among populations than do nuclear markers. This increased divergence may be a result of the smaller effective population size of mitochondrial DNA than the nuclear genome, which makes it more susceptible to drift (Billington 2003). In addition, many species demonstrate positive selection in the mitochondrial genome, which could further increase divergence among populations[1]. Finally, phenotypic data such as meristic traits (Nolte and Sheets 2005) or isotope data (Rundel et al. 2013) could also be evaluated in the same manner as haploid data within the framework of MSA.

Here we evaluated the precision and accuracy of a developing chum salmon baseline with a method that we call 'leave-ten-percent-out cross-validation' (LTO) as an alternative to ONCOR. LTO is derived from the K-fold cross-validation method of classification statistics (Geisser 1975, Hastie et al. 2001) and has been recommended for use in MSA applications for fisheries (Waples 2010). Our LTO method reduces optimism of baseline evaluation,

[1]M. R. Garvin, School of Fisheries and Ocean Sciences, University of Alaska Fairbanks, unpublished data.

accommodates haploid and diploid data, and the MSAs can be estimated by either Bayesian or ML methods. Here we limit consideration to Bayesian methods of estimation because they provide some advantages to ML methods and given that they are frequently used in practice, it would be more consistent to evaluate a baseline with the same method that will be used when it is put into practice. Computer programs to perform Bayesian MSA include BAYES (Pella and Masuda 2001) and cBayes (Neaves et al. 2005) although cBayes does not currently accept haploid data. We provide code in R that generates the input files for the computer program BAYES (Appendix 2).

Many baseline development projects are concerned with conversion from microsatellite-based baselines to SNP-based baselines. Attempts to quantify the number of SNPs that provide equivalent discrimination to microsatellite loci often compare either equivalent numbers of alleles (Kalinowski 2002) or the number of SNPs and microsatellites surveyed on the same individuals (Narum et al. 2008, Santure et al. 2010). As an alternative, we simulated increasing numbers of SNP loci from the current baseline data and evaluated the anticipated baseline performance from the combined laboratory-derived loci with LTO. We then used logistic regression to extrapolate the number of SNPs necessary to provide a given degree of accuracy. Key to the extrapolation is an assumption that the same methods used in discovery of the current SNP baseline will be used to search for the future informative SNP loci.

MATERIALS AND METHODS

Generation of Test Baselines and Stock Mixtures

Samples from 74 chum salmon stocks taken from populations across the Pacific Rim were obtained between 1987 and 2008 (Fig. 1; Table 1). Each individual in the baseline[2] was genotyped with nine microsatellite loci and 23 SNPs representing 12 SNP loci (Table 2). The SNP loci were developed to maximize divergence among stocks (see Appendix 1) and the stocks were sorted into 25 reporting groups based on geography and management areas delineated in previous work (e.g., Beacham et al. 2009, Seeb et al. 2011) as well as from our principal components analysis (PCA) of the current baseline (Appendix Fig. 1-1). Additional loci and additional stocks will be added to the baseline as management problems are identified and funds become available.

We divided each baseline stock sample sequentially into 10 equal parts with code in R (Appendix 2). The individuals that composed one part from each stock were combined into a test mixture of 450 individuals, and the genetic information from the remaining nine parts was used as the baseline data in the mixture analyses performed with the program BAYES (Pella and Masuda 2001). The BAYES analyses were repeated 10 times; each time, a different one of the 10 parts was used as the test mixture and the remaining 9 parts served as the baseline. The LTO method guarantees that each individual is used in a test mixture once and in a test baseline nine times, except for 'remainder' individuals (see below), and that no individual occurs in both a test mixture and the test baseline used in its analysis (no bi-presence). The stock compositions of all test mixtures were identical.

[2]Data are available upon request from the lead author.

In our baseline, the sample sizes of many stocks differed and many were not evenly divisible by 10 (Table 1). Therefore, the composition of the mixture was nearly proportional to baseline sample sizes. In addition, when we divided our baseline stock samples into 10 equal parts, some individuals remained. These 'remainder' individuals were added to each of the 10 test baselines but not included in any of the test mixtures. The term 'test dataset' will be used to denote one of the 10 'test mixtures' and the associated 'test baseline'.

Measurements of Diversity

Many of the loci that we evaluated in this work are potentially informative for MSA because they demonstrate large divergence estimates among baseline populations (Storer et al. 2012). We used GDA (Lewis and Zaykin 2001) to calculate F_{ST} (Weir and Cockerham 1984) locus by locus and overall, ϕ_{ST} (the haploid equivalent of F_{ST}), expected heterozygosity (H_e) for nuclear loci and haplotype diversity for the mitochondrial locus (Excoffier et al. 1992). We used the 'adegenet' package (Jombart 2008) to calculate D_{EST} - Jost's D (Jost 2008) (methods for the haploid data were identical).

As noted earlier, some optimism could have been introduced into our baseline evaluations from error-enhanced divergence (Anderson 2010). However, for the baseline samples, BAYES revises the observed allele frequencies by objectively shrinking them toward a better-determined central value among stocks (Pella and Masuda 2001). To demonstrate the shrinkage effects, we compared apparent diversity before and after the revision of the allele frequencies. We calculated G_{ST} values for each locus with the methods of Nei and Chesser (1983) with the observed allele frequencies from the entire baseline and then for each of the 10 baselines that were created during LTO. We then performed the same calculations with the mean

allele frequencies of the baseline posterior, computed as a weighted average of the original frequencies and the prior grand mean (Equation 4 in Pella and Masuda 2001). This latter baseline posterior distribution serves as the prior for the analysis of the mixture sample.

BAYES LTO

We used the program BAYES to estimate the proportion that each of the 74 stocks contributed to each test mixture. The low information Dirichlet prior probability distribution for stock proportions (its weight during estimation counts as a single individual added to the mixture of 450 individuals) was set to equal proportions among the entire 74 stocks. Three independent MCMC chains were run for each test mixture with different starting values for stock proportions.

The three MCMC chains were run with the BAYES program to obtain 400,000 samples of the unknowns (stock proportions and baseline genetic parameters) from their posterior distribution and every 40^{th} sample was saved. The first half of each chain was discarded as burn-in to remove dependence on starting values. The second halves of the three chains were combined to provide a total sample of 15,000 stock composition estimates from their posterior distribution. Gelman and Rubin statistics were computed for each of the 25 reporting groups (all were less than 1.2) to verify that pooled samples from the three chains had converged to the posterior distribution for the regional composition (Gelman and Rubin 1992).

The results from BAYES for the 10 test datasets provided a sample of 10 posterior distributions of stock composition estimates for the mixtures of the 74 baseline stocks. The posterior average for each of these distributions was considered the point estimate. Regional compositions were obtained as sums over the point estimates for the individual stocks of the regional groups; and their means and variances, as well as the observed bias and mean square

errors from the true and known regional proportions, were computed. After the BAYES analysis, we assigned each individual in the mixture sample to the source stock with the highest posterior probability; that is, the so-called maximum *a posteriori* (MAP) rule.

For comparative purposes, we generated stock composition estimates with the program SPAM with the same 10 test datasets. We also generated estimates with SPAM in simulation mode with the same 10 test baselines that were used in the BAYES and SPAM LTO, but we simulated mixtures that had the expected composition to the mixtures created with LTO (Appendix Figs. 1-2 to 1-7).

How Many SNPs Would Equal the Discriminatory Power of the Current Combined SNP and Microsatellite Baseline?

To simulate "new" hypothetical nuclear SNP loci, we used the same data from 74 chum salmon populations genotyped with 23 SNPs, which represented 12 loci (11 nuclear loci and the mitochondrial genome). These new nuclear loci were obtained by randomly drawing, with replacement, additional loci from the 11 empirically genotyped nuclear SNP loci for which real data were available. That is, we assumed the 11 SNP loci found by our laboratory methods represented a random sample from a large population of independent SNPs for chum salmon. We did not include the mitochondrial data in the simulation of the new loci because the genetic material in the mitochondrion of an individual behaves as a single locus so generation of additional mitochondrial loci would not be biologically meaningful.

A locus was randomly drawn from the 11 nuclear SNP loci and corresponding single-locus genotypes were generated for each original baseline individual by drawing a pair of alleles without replacement between the two draws from the sample of its SNP baseline stock. Multi-locus SNP genotypes for individuals were generated by parallel independent sampling for

additional new loci, followed by concatenation of the outcomes. This strategy preserves the genetic information from our original informative loci but adds the multi-locus variability that would be expected at additional unlinked loci. The new SNP loci were appended to the original 11 SNP loci and the mitochondrial locus to generate seven extended SNP sets with 20, 30, 40, 50, 60, 70, and 80 SNP loci. For example, 8 new SNP loci were simulated and added to the 11 original SNP loci and the mitochondrial locus to create 20 locus genotypes. The baseline samples were then divided into 10 equal test datasets and analyzed with BAYES LTO.

We assigned each of the 450 individuals in a test mixture to the stock with the highest posterior probability and thereby to the reporting group of origin. The proportion of individuals correctly assigned to a reporting group was calculated for each test mixture and the corresponding average of the 10 test mixtures from the LTO was then calculated. For instances in which identical posterior source probabilities were calculated for an individual belonging to either of two reporting groups, we randomly assigned the individual to one of the groups. The number of identical probabilities that occurred was inconsequential and ranged from zero to five for a single test mixture. These equivocal individuals accounted for only 0.3% of the total assignments for the 10 test mixtures and their frequency of occurrence did not appear to be related to the number of loci.

To determine the number of SNP loci required for a high degree of correct assignment, we used logistic regression to estimate the correct proportional assignment to reporting group (θ) for a given number of SNPs (X) based on the empirical performance of the extended SNP sets for 20 to 80 SNPs. The log odds regression model relating θ and X is

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 X_i,$$

where θ_i is the probability of correct assignment of a random individual to its reporting group, X_i is the number of SNP loci ($X_1 = 20$, $X_2 = 30$,.., $X_7 = 80$), β_0 is the intercept parameter, and β_1 is the slope. We fit the model to the counts of correctly and incorrectly assigned individuals as related to the number of SNPs by the ML method. The fit was performed in each of the 10 cross-validation datasets which provided estimates of β_0 and β_1. Finally, the estimates for those parameters were used to extrapolate an estimated number of SNPs necessary to achieve 90% correct assignment ($X^{90\%}$) for each of the 10 datasets.

RESULTS

Evaluation of Combined Microsatellite and SNP Baseline

We estimated the stock compositions and their standard errors for each of the 10 test datasets with the program BAYES (Appendix Table 1-1; Fig. 2a). Three regional groups stand out for their higher standard errors: the Lower Yukon River, the Lower Kuskokwim River, and Behm Canal. We based our regional groups on what has already been reported in the literature (Beacham et al. 2009, Seeb et al. 2011) as well as our PCA (Appendix Fig. 1-1). We also display the baseline performance as in Anderson et al. (2008), with the mean stock group composition estimates paired with their true values and a reference line that indicates perfect accuracy (Fig. 2b). For comparison, LTO results were obtained with the program SPAM with the ML method and with SPAM in simulation mode but with 10 simulated mixtures with expected composition matching the composition of the LTO mixtures (Appendix Figs. 1-2 to 1-4).

Evaluation of the Bias and Accuracy of Estimates with BAYES

We quantified the overall accuracy of BAYES LTO by calculating the absolute difference between the mean of the 10 test estimates of a regional proportion and the true value for each of the 25 reporting groups (Fig. 3). A few regions had relatively high estimates of absolute bias of a mixture proportion from 0.01 to 0.02.

We evaluated the apparent precision of BAYES LTO by calculating the standard deviation of the 10 estimates for each of the 25 reporting groups (Fig. 4). The 10 estimates for any method are not independent because baseline and mixture samples among sets overlap to some degree, which is why we say apparent precision. Many of the reporting groups that had high estimated biases also had high standard deviations.

Finally, we combined the bias and variance estimates into the mean squared error (we report the square root to provide equivalent measures to the accuracy and precision estimates) (Fig. 5). For comparative purposes, we also include these three estimates when the program SPAM was used for LTO and in simulation mode (Appendix Figs. 1-5 to 1-7). As was expected, the simulation mode of SPAM provides optimistic assessments with both low bias and high precision. For regions in western Alaska that are difficult to delineate, SPAM LTO demonstrated higher bias but higher precision compared to BAYES. That is, the estimate was not as accurate but appeared to be a better estimate because the precision was higher.

Evaluation of Potential Bias in Baseline

Optimism could have been introduced into our baseline evaluation because we used some of the same samples to choose SNPs and to evaluate them (high-grading bias). However, the subset of samples used both to choose and to evaluate the SNPs was small and our large sample

sizes for the ascertainment panels likely offset some bias (see Appendix 1). Further optimism is possible through error-enhanced divergence from baseline sampling (Anderson et al. 2008). In order to determine how inflated divergence estimates might affect our baseline evaluation, we compared locus-by-locus G_{ST} values for the original baseline with corresponding mean G_{ST} values computed for the 10 BAYES LTO baselines (Fig. 6). For some loci, the G_{ST} values of the LTO baselines were indeed inflated, which suggests that optimism would be introduced from the smaller sample sizes that resulted when we created the mixtures. However, the locus-by-locus G_{ST} values for the 10 BAYES LTO baselines with allele frequencies that were shrunk toward their prior means from Equation 4 in Pella and Masuda (2001) were all smaller than for the original allele frequencies (Fig. 6). Therefore, the shrinkage directly opposes optimism in the evaluation from error-enhanced divergence.

Future Baseline Improvements

We plotted the assignment of each individual across all geographic groups[3] to identify: 1) if they were assigned correctly to their population of origin, 2) if they were assigned incorrectly to the population of origin but correctly to their geographic reporting group, or 3) to which population they were incorrectly assigned (Fig. 7). We chose to display three examples: one that mis-assigns individuals to a population of origin (true population is Sopochnoe) but correctly to region (Kurils), one that mis-assigns individuals to both population and region (true population and region are Anadyr), and one that assigns individuals correctly to population (Big Creek) and region (S. Bristol Bay).

[3]Data for all 74 populations are available upon request from the lead author.

How Many SNP Loci are Needed to Exceed or Equal the Combined SNP and Microsatellite Baseline?

We calculated the mean proportion of individuals correctly assigned by the MAP rule to the 25 reporting groups from the 10 test datasets. These values were determined for each of the 20, 30, 40, 50, 60, 70, and 80 extended SNP locus genotypes ('Simulated Loci', Fig. 8) and for the original combined microsatellite and SNP baseline (horizontal dashed line, Fig. 8). A reference line is provided for a hypothetical 90% correct assignment accuracy for all individuals in the baseline (horizontal solid line). Between 50 and 60 informative SNP loci appear necessary to equal the combined SNP and microsatellite baseline.

To identify further improvement if more SNP loci were used we fit an asymptotic curve with logistic regression to the simulated data for each of the 10 extended datasets with 20, 30, 40, 50, 60, 70, and 80 SNP loci. The mean values for the slope and intercept parameters from these 10 fitted models were used to display the relationship between the number of SNPs and the proportion of correctly assigned individuals (Fig. 8). We estimated the number of SNP loci needed to achieve 90% accuracy (a common value used in MSA) to region from the fitted curve of each extended dataset, which averaged 125 SNPs with a standard error of $\pm$5.4. This number corresponds to the mean percent correct assignment over all reporting groups. Specific reporting groups may require more or fewer SNPs depending on the divergence among those stocks.

DISCUSSION

Mixed-stock analysis is an important tool for the conservation and management of numerous species, and technological advances will likely continue to provide a growing wealth of genetic data as well as the ability to generate millions of genotypes rapidly and inexpensively

(Larson-Cook et al. 2011, Ragoussis 2009). As a result, MSA will increasingly be used to manage and monitor an ever-wider array of species and stocks. These applications of MSA will require trustworthy baseline evaluations in order to meet the challenges to managers from the affected resource users and stakeholders.

Our results demonstrate how the LTO method can be used to evaluate a baseline for any species and identify where future resources should be devoted. For example, the output of our analysis indicates that individuals from populations in coastal western Alaska are difficult to assign correctly to their population or even geographic region of origin. Consequently, for this baseline and for this species, informative SNP markers should be developed to improve assignment to that geographical region. Output from LTO can be combined with other known information to further direct marker development efforts. For instance, the difficulty of assignment of chum salmon (*Oncorhynchus keta*) from this geographic area may reflect their recent colonization after the Last Glacial Maximum (Seeb and Crane 1999, Wilmot et al. 1994), high levels of gene flow (Olsen et al. 2010), or both (Garvin et al. 2013) and therefore markers that are responsible for local adaptation and that show higher levels of genetic divergence than neutral markers may be necessary to improve MSA. New markers could be developed and then re-evaluated with LTO.

Our results also emphasize several advantages compared to simulation-based evaluations. First, the LTO method can use Bayesian or ML methods for MSA while the simulation-based method implemented in ONCOR is currently limited to a modified ML algorithm. We use Bayesian MSA here because it is frequently used in applications now and the baseline evaluation should employ the same estimation technique for accuracy and consistency. The use of Bayesian methods also adds scope to the types of data and models that can be used for MSA. Haploid

genetic data can now be included in baseline evaluations using BAYES LTO, which is highly informative for many species including chum salmon (see Appendix Fig. 1-1). Importantly, spatial, temporal, and morphometric data for individuals could also be used with this method in the future to increase the ability to distinguish among stock sources (Nolte and Sheets 2005, Barbee and Swearer 2007, Gomez-Diaz and Gonzales-Solis 2007, Reich et al. 2008, Reich and Bondell 2011).

Second, the LTO method uses real multi-locus genotypes observed in the baseline samples for creating the mixtures, which adds authenticity to the evaluation that is easily understood. The simulation-based methods always include genetic assumptions for genotype frequencies in the separate stocks in order to generate the hypothetical genotypes of the mixture individuals, and these assumptions could reduce trust in the baseline evaluations.

Third and last, each individual from the baseline (except the few remainder individuals) occurs in one test mixture for the LTO method. Therefore, the assignments of nearly all baseline individuals are known, which can provide useful information that is unavailable with simulation-based methods. For example, an individual may be mis-assigned because of missing data from multiple loci, and that individual could be easily identified by a search of the database. Also, the stock source and destination of mis-assigned individuals can be used to form regional groups for which estimates will have lower bias and higher precision. For this LTO method, the individual assignment not only indicates which stock or region of interest is difficult to correctly assign but specifically the stock or region to which individuals are mis-assigned. This provides information on the geographic regions for which future baseline development efforts should be directed.

Several baseline evaluations reported in the literature use ‘100% proof tests’ to evaluate genetic baselines (Habicht et al. 2010, Seeb et al. 2011, Templin et al. 2011) in which several

hundred individuals are removed from the baseline as a test mixture to be evaluated with the now reduced baseline. Proof tests avoid the optimism discussed by Anderson et al. (2008), can accommodate haploid data and use Bayesian-based methods, but require a sufficient number of samples to accommodate the removal of a few hundred individuals to produce the mixtures. If species have small census sizes or if only a portion of the baseline is to be evaluated, then sample sizes will be insufficient.

Furthermore, the proof tests have been used with 100% mixtures instead of the nearer-to-equal contribution mixtures with our LTO method (this is also often the case with simulation methods; for example, Beacham et al. 2009, Seeb et al. 2011). These 100% mixtures inadvertently introduce further optimism into baseline evaluation because their composition is easier to determine. The accuracy and precision of estimated mixture composition is limited by success in assignment of individuals to the baseline stocks. Highest accuracy and precision occur when individuals of all stocks can be correctly assigned and the simple multinomial sampling model applies. As percent correct assignments decline, sampling variation and bias accrue in the mixture analyses (Pella and Robertson 1979). Whether ML or Bayesian estimation is used, assignments are based on the posterior source probabilities of the mixture individuals (see Equation 5 of Pella and Milner 1987). When the true compositions are used in the posterior source probabilities, the mixture individuals can be assigned with certainty to their source for 100% mixtures, but not for the equal proportions mixtures. The estimated mixture proportions will vary and approximate these true values as the estimation algorithms cycle; consequently, correct classification will be less for the equal-proportions mixture and higher for the 100% mixture.

The main limitations of LTO are that the number and possible mixtures to include in the baseline evaluation are few. The simulation-based methods generate hypothetical genotypes for individuals in these mixtures of arbitrary stock composition and size by computer modeling of the baseline data. The number of times that a mixture of a specified stock composition and size can be generated and analyzed is arbitrarily large (typically 100 to 1,000 times) but this is not possible with our LTO method because Bayesian estimation is computationally intensive.

The development of genetic baselines can be costly in both time and resources if the number of markers needed to provide accurate composition estimates of mixtures or assignments of individuals is unknown prior to baseline development. Open-ended development of genetic baselines has been the standard practice, and the number of SNPs needed has depended on the target species of interest, the amount of divergence among stocks, and the stocks that were included in the analysis (Campbell et al. 2012, Campbell and Narum 2008, Elfstrom et al. 2006, Griffiths et al. 2010, Smith et al. 2005). We provide a method to estimate the number of genetic markers that will be needed for MSA of a specified accuracy given an initial small set of *informative* markers. We showed that, for this chum salmon baseline, about 60 *informative* SNPs would be equivalent to the nine microsatellite and 12 *informative* SNP loci in the baseline. Furthermore, a logistic regression analysis predicted that a baseline with about 125 *informative* SNPs found by our laboratory techniques would be needed to assign stocks with at least 90% accuracy to a group. Our cross-validation method provides managers and scientists with the ability to place a direct cost on the development of accurate baselines.

ACKNOWLEDGMENTS

We would like to thank the Rasmuson Foundation, the Arctic-Yukon-Kuskokwim Sustainable Salmon Initiative (www.aykssi.org) awarded through the Bering Sea Fishermen's Association, the University of Alaska Experimental Program to Stimulate Competitive Research (EPSCoR), and the National Oceanic and Atmospheric Administration's (NOAA) Alaska Fisheries Science Center (AFSC) for funding this work. This study was also supported by a grant of the HPC resources from the Arctic Region Supercomputing Center. The findings and conclusions presented by the authors, however, are their own and do not necessarily reflect the views or positions of the reviewers, funding agencies, the University of Alaska Fairbanks, School of Fisheries and Ocean Sciences, or the National Marine Fisheries Service, NOAA.

CITATIONS

Anderson, E. 2010. Assessing the power of informative subsets of loci for population assignment: Standard methods are upwardly biased. Mol. Ecol. 10:701–710.

Anderson, E., R. S. Waples, and S. T. Kalinowski. 2008. An improved method for estimating the accuracy of genetic stock identification. Can. J. Fish. Aquat. Sci. 65:1475–1486.

Barbee, C., and S. E. Swearer. 2007. Characterizing natal source population signatures in the diadromous fish *Galaxias maculatus*, using embryonic otolith chemistry. Mar. Ecol. Progr. Ser. 343:273–282.

Beacham, T. D., J. R. Candy, and C. Wallace. 2009. Microsatellite stock identification of chum salmon on a Pacific Rim basis. N. Am. J. Fish. Manage. 29:1757–1776.

Billington, N. 2003. Mitochondrial DNA, p. 59–100. *In* E. Hallerman (editor), Population Genetics: Principles and Applications for Fisheries Scientists. American Fisheries Society, Bethesda.

Bowen, B., W. S. Grant, Z. Hillis-Starr, D. J. Shaver, K. A. Bjorndal, A. B. Bolten, and A. L. Bass. 2007. Mixed-stock analysis reveals the migrations of juvenile hawksbill turtles (*Eretmochelys imbricata*) in the Caribbean Sea. Mol. Ecol. 16:49–60.

Campbell, N., S. J. Amish, V. L. Pritchard, K. M. McKelvey, M. K. Young, M. K. Schwartz, J. C. Garza, G. Luikart, and S. R. Narum. 2012. Development and evaluation of 200 novel SNP assays for population genetic studies of westslope cutthroat trout and genetic identification of related taxa. Mol. Ecol. Res. 12:942–949.

Campbell, N. R., and S. R. Narum. 2008. Identification of novel single-nucleotide polymorphisms in chinook salmon and variation among life history types. Trans. Am. Fish. Soc. 137:96–106.

Debevec, E. M., R. B. Gates, M. Masuda, J. Pella, J. Reynolds, and L. W. Seeb. 2000. SPAM (version 3.2): Statistics Program for Analyzing Mixtures. J. Heredity 91:509–510.

Elfstrom, C., C. Smith, and J. Seeb. 2006. Thirty-two single nucleotide polymorphism markers for high-throughput genotyping of sockeye salmon. Mol. Ecol. Notes 6:1255–1259.

Excoffier, L., P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479–491.

Garvin, M. R., and A. J. Gharrett. 2007. DEco-TILLING: an inexpensive method for SNP discovery that reduces ascertainment bias. Mol. Ecol. Notes 7:735–746.

Garvin, M. R., and A. J. Gharrett. 2010. Application of SNP markers to chum salmon (*Oncorhynchus keta*): Discovery, genotyping, and linkage phase resolution. J. Fish Biol. 77:2137–2162.

Garvin, M. R., K. Saitoh, V. Brykov, D. Churikov, and A. J. Gharrett. 2010. Single nucleotide polymorphisms in chum salmon (*Oncorhynchus keta*) mitochondrial DNA derived from restriction site haplotype information. Genome 53:501–507.

Garvin, M., C. M. Kondzela, P. C. Martin, J. Guyon, W. D. Templin, T. Dann, N. DeCovich, S. E. Gilk, and A. J. Gharrett. 2013. Recent physical connections among now divided drainages may explain weak genetic structure in western Alaskan chum salmon (*Oncorhynchus keta*) populations. Ecol. Evol. 3:2362–2377.

Geisser, S. 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70:320–328.

Gelman, A., and D. B. Rubin. 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7:457–511.

Gomez-Diaz, E., and J. Gonzales-Solis. 2007. Geographic assignment of seabirds to their origin: Combining morphological, genetic, and biogeochemical analyes. Ecol. Appl. 17:1484–1498.

Griffiths, A. M., G. Machado-Schiaffino, E. Dillane, J. Coughlan, J. L. Horreo, A. E. Bowkett, P. Minting, S. Toms, W. Roche, P. Gargan, P. McGinnity, T. Cross, D. Bright, E. Garcia-Vazquez, and J. R. Stevens. 2010. Genetic stock identification of Atlantic salmon (*Salmo salar*) populations in the southern part of the European range. BMC Genetics 11:31.

Habicht, C., L. W. Seeb, K. W. Myers, E. V. Farley, and J. E. Seeb. 2010. Summer–fall distribution of stocks of immature sockeye salmon in the Bering Sea as revealed by single-nucleotide polymorphisms. Trans. Am. Fish. Soc. 139:1171–1191.

Hastie, T., R. Tibshirani, and J. Friedman. 2001. Elements of statistical learning: Data mining, inference and prediction. Springer-Verlag, New York, NY. 533 p.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405.

Jost, L. 2008. G_{ST} and its relatives do not measure differentiation. Mol. Ecol. 17:4015–4026.

Kalinowski, S. T. 2002. How many alleles per locus should be used to estimate genetic distances? Heredity 88:62–65.

Kalinowski, S. T. 2003. Genetic Mixture Analysis 1. Department of Ecology, Montana State University. http://www.montana.edu/kalinowski/GMA/GMA_Home.htm.

Larson-Cook, K., E. V. Zon, S. Rai, and T. Rusch. 2011. Microplate replacement for HT screening: Technology miniaturizes reactions in highly automated inline platform. Genetic Eng. Biotech. News 31:40–41.

Lewis, P. O., and D. Zaykin. 2001. Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). http:lewis.eeb.uconn.edu/lewishome/software.html.

Manel, S., O. E. Gaggiotti, and R. S. Waples. 2005. Assignment methods: Matching biological questions with appropriate techniques. Trends Ecol. Evol. 30:136–142.

Moriya, S., S. Shunpei, T. Azumaya, O. Suzuki, and S. Urawa. 2006. Genetic stock identification of chum salmon in the Bering Sea and North Pacific Ocean using mitochondrial DNA microarray. Mar. Biotech. 9:179–191.

Narum, S. R., M. Banks, T. D. Beacham, M. R. Bellinger, M. R. Campbell, J. Dekoning, A. Elz, C. M. Guthrie, C. Kozfkay, K. M. Miller, P. Moran, R. Phillips, L. W. Seeb, C. T. Smith, K. Warheit, S. F. Young, and J. C. Garza. 2008. Differentiating salmon populations at broad and fine geographic scales with microsatellites and single nucleotide polymorphisms. Mol. Ecol. 17:3464–3477.

Neaves, P., C. G. Wallace, J. R. Candy, and T. D. Beacham. 2005. cBayes: Computer program for mixed stock analysis of allelic data. Version v5.01. http://www.pac.dfo-mpo.gc.ca/sci/mgl/Cbayes_e.htm.

Negrini, R., L. Nicoloso, P. Crepaldi, E. Milanesi, L. Colli, F. Chegdani, L. Pariset, S. Dunner, H. Leveziel, J. L. Williams, and P. Ajmone Marsan. 2008. Assessing SNP markers for assigning individuals to cattle populations. Anim. Genet. 40:18–26.

Nei, M., and R. K. Chesser. 1983. Estimation of fixation indices and gene diversities. Annal. Hum. Genet. 47:253–259.

Nolte, A.W., and H. D. Sheets. 2005. Shape based assignment tests suggest transgressive phenotypes in natural sculpin hybrids (Teleostei, Scorpaeniformes, Cottidae). Front. Zool. 2:11.

Olsen, J. B., P. A. Crane, B. G. Flannery, K. Dunmall, W. D. Templin, and J. K. Wenburg. 2010. Comparative landscape genetic analysis of three Pacific salmon species from subarctic North America. Conserv. Genetics 12:223–241.

Pella, J., and M. Masuda. 2001. Bayesian methods for analysis of stock mixtures from genetic markers. Fish. Bull., U.S. 99:151–167.

Pella, J., and G. Milner. 1987. Use of genetic markers in stock composition analysis, p. 247–275. *In* N. Ryman and F. Utter (editors), Population Genetics and Fishery Management. Washington Sea Grant Program, Univ. Wash. Press, Seattle, WA.

Pella, J., and T. L. Robertson. 1979. Assessment of composition of stock mixtures. Fish. Bull., U.S. 77:387–398.

Ragoussis, J. 2009. Genotyping technologies for genetic research. Ann. Rev. Human Genetics 10:117–133.

Reich, B. J., and H. D. Bondell. 2011. A spatial Dirichlet process mixture model for clustering population genetics data. Biometrics 67:381–390.

Reich, D., A. L. Price, and N. Patterson. 2008. Principal component analysis of genetic data. Nat. Genetics 40:491–492.

Rundel, C. W., M. B. Wunder, A. H. Alvarado, K. C. Ruegg, R. Harrigan, A. Schuh, J. F. Kelly, R. B. Siegel, D. F. DeSante, T. B. Smith, and J. Novembre. 2013. Novel statistical methods for integrating genetic and stable isotope data to infer individual-level migratory connectivity. Mol. Ecol. 22:4163–4176.

Santure, A. W., J. Stapley, A. D. Ball, T. R. Birkhead, T. Burke, and J. Slate. 2010. On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Mol. Ecol. 19:1439–1451.

Seeb, L. W., and P. A. Crane. 1999. Genetic heterogeneity in chum salmon in western Alaska, the contact zone between northern and southern lineages. Trans. Am. Fish. Soc. 128:58–87.

Seeb, L. W., W. D. Templin, S. Sato, S. Abe, K. Warheit, J. Y. Park, and J. E. Seeb. 2011. Single nucleotide polymorphisms across a species' range: Implications for conservation studies of Pacific salmon. Mol. Ecol. Res. 11:195–217.

Smith, C. T., C. M. Elfstrom, L. W. Seeb, and J. E. Seeb. 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol. Ecol. 14:4193–4203.

Sokal, R. R., and F. J. Rohlf. 1994. Biometry: The principles and practices of statistics in biological research. W. H. Freeman, New York, NY. 880 p.

Storer, C., C. E. Pascal, S. B. Roberts, W. D. Templin, L. W. Seeb, and J. E. Seeb. 2012. Rank and order: Evaluating the performance of SNPs for individual assignment in a non-model organism. PLoS ONE 7:e49018.

Templin, W. D., J. E. Seeb, J. R. Jasper, A. W. Barclay, and L. W. Seeb. 2011. Genetic differentiation of Alaska Chinook salmon: the missing link for migratory studies. Mol. Ecol. Res. 11:226–246.

Waples, R. S. 2010. High-grading bias: subtle problems with assessing power of selected subsets of loci for population assignment. Mol. Ecol. 19:2599–2601.

Wasser, S., A. M. Shedlock, K. Comstock, E. A. Ostrander, B. Mutayoba, and S. Mathew. 2004. Assigning African elephant DNA to geographic region of origin: Applications to the ivory trade. Proc. Nat. Acad. Sci. USA 101:14847–14852.

Weir, B., and C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1360.

Wilmot, R. L., R. J. Everett, W. J. Spearman, R. Baccus, N. V. Varnavskaya, and S. V. Putivkin. 1994. Genetic stock structure of western Alaskan chum salmon and a comparison with Russian far east stocks. Can. J. Fish. Aquat. Sci. 51(Suppl. 1):84–94.

Table 1. -- Stocks used in the study and their sources*. The regional groupings and the baseline sample sizes are provided.

Stock No.	Stock name	Date	Latitude	Longitude	25 Group No.	25 Group name	Sample size	Source
1	Tsugarishi	1991	39.20	141.80	1	Honshu	40	KITU
2	Katagishi	1991	39.60	142.00			40	KITU
3	Reidovaya	2006	45.38	147.98	2	Kurils	30	LGG
4	Sopochnoe Lake	2004	45.32	148.41			48	LGG
5	Naiba	1995/1996	47.45	142.76	3	S. Sakhalin	47	RAS
6	Okhotsk	2003	46.87	143.17			23	RAS
7	Taranai	2003	46.63	142.43			25	LGG
8	Udararnitsa	1994	46.80	143.30			48	RAS
9	Tym	1995/2003	51.26	142.71	4	N. Sakhalin	71	RAS/ABL
10	Heilong	1994	48.38	134.38	5	Amur	44	X.Luan
11	Amur Early	2003	52.93	141.17			45	RAS
12	Amur Late	2003	52.93	141.17			27	RAS
13	Anuyi	2002	49.32	136.47	6	Primore	46	RAS
14	Barabashevka	1994/1995	43.11	131.64			50	RAS
15	Narva	1995/2005	42.99	131.49			67	RAS
16	Ryazakanovka	1994/1995	43.16	132.11			63	RAS
17	Suifen	1994	43.34	131.82			20	RAS
18	Ola	1999	59.60	151.27	7	Magadan	43	RAS
19	Taui	1999	59.39	149.14			37	RAS
20	Hailula	2003	58.20	162.03	8	Kamchatka	47	RAS
21	Ossoro	1996	59.18	163.15			48	TNRO
22	Hairsova	1990/1993	57.09	156.52			96	RAS
23	Kol	2003	53.81	155.94			47	RAS
24	Utka	2002	53.15	156.08			35	RAS
25	Oclan	1993	62.77	164.33	9	Anadyr	73	RAS
26	Anadyr	1991	64.90	176.22			111	RAS
27	Kanchalon	1991	65.12	176.53			79	ABL
28	Salcha	1994	64.47	-146.98	10	Middle Yukon	96	ADF&G
29	Toklat	1994	64.45	-150.31			96	ADF&G

Table 1. -- Continued.

30	FishBranch	1992	66.45	-138.58	11	Upper Yukon	96	ADF&G
31	Kluane	1992	61.88	-139.72			96	USFWS
32	Sheenjek	1988/1989	66.74	-144.57			96	ADF&G
33	Teslin	1992	61.57	-134.90			96	USFWS
34	Kobuk	2000	66.92	-160.81	12	Lower Yukon	96	ADF&G
35	Agiapuk		65.17	-165.68			96	DFO
36	Pilgrim	2004	65.16	-165.22			96	KWRK
37	Snake	2004	64.50	-165.41			96	KWRK
38	Pikmitalik	2004	63.27	-162.60			96	KWRK
39	Atchulingak	1989	61.96	-162.83			96	USFWS
40	Anvik	1989	62.68	-160.20			75	USFWS
41	Kaltag	1992	64.33	-158.72			48	USFWS
42	Nulato	2003	64.71	-158.14			48	USFWS
43	Kanektok	1989	59.75	-161.93	13	Lower Kuskokwim	75	ADF&G
44	Kasigluk	1990	60.85	-161.23			73	ADF&G
45	Kwethluk	1989	60.81	-161.45			77	ADF&G
46	Goodnews	1989	59.13	-161.48			96	ADF&G
47	Nushagak	1988	58.80	-158.63			75	ADF&G
48	Bigcreek	1988/2000	58.29	-157.53	14	S. Bristol Bay	96	ADF&G
49	Gertrude	1987/1999	58.17	-156.21			96	ADF&G
50	Meshik	1989	56.81	-158.66			75	ADF&G
51	Frosty	2000	55.07	-162.81	15	Frosty	96	ADF&G
52	Kizyuak	1989	57.82	-152.80	16	Kodiak	48	ADF&G
53	Little Susitna	1990	61.25	-150.29	17	Cook Inlet	39	ABL
54	Olsen	1992/1997	60.76	-146.17	18	Prince William Sound	96	ABL
55	Alsek	2000	59.13	-138.62	19	Yakutat	96	ABL
56	EAlsek	2006	59.11	-138.52			48	UAFSFOS
57	Green's Creek	1995	58.10	-134.76	20	Northern SE Alaska	96	ABL
58	Herman Creek	1987/1990/2008	59.42	-136.10			96	ABL
59	Taku	2000	58.43	-133.98			45	ABL
60	Blossom	1986	55.40	-130.61	21	Behm Canal	48	ABL

Table 1. -- Continued.

61	Marten	1986	55.16	-130.53			48	ABL
62	Portage Creek	1986/1988	55.77	-131.04			96	ABL
63	Wilson	1986	55.40	-130.61			40	ABL
64	Herman River	1986	55.99	-131.27			40	ABL
65	Karta	1986	55.56	-132.57	22	Prince of Wales Island	48	ABL
66	Old Tom Creek	1986/1988	55.40	-132.40			96	ABL
67	Bag Harbor	1989	52.35	-131.36	23	QCI	48	ABL
68	Tasu	1989	52.87	-132.08			48	ABL
69	Klownick	1989	52.38	-126.75	24	N. British Columbia	48	ABL
70	Neekas	1989	52.47	-128.17			48	ABL
71	Grant	1998	48.27	-122.02	25	Puget Sound	96	WDFG
72	Kennedy	1996	47.10	-123.09			96	WDFG
73	Johns	2003	47.24	-123.04			96	WDFG
74	Quilcene	1997	47.80	-122.86			40	ABL

*ADF&G – Alaska Department of Fish & Game, DFO – Department of Fisheries Oceans, Canada, KWRK – Kawerek, LGG – Laboratory of Genetic Identification, Institute of General Genetics, ABL – Auke Bay Laboratories, Alaska Fisheries Science Center, TNRO – Kamchatka TINRO, UAFSFOS – University of Alaska Fairbanks, USFWS – U.S. Fish and Wildlife Service, WDFG– Washington Department of Fish and Game, RAS – Russian Academy of Sciences, KITU - Kitasato University.

Table 2. -- Measures of genetic diversity for the 12 single nucleotide polymorphism (SNP) and nine microsatellite loci analyzed for all individuals in the 74 chum salmon stocks in this work. F_{ST} is Weir and Cockerham's θ (Weir and Cockerham 1984), D_{EST} is Jost's D (Jost 2008), and H_e is the expected heterozygosity. Some SNP loci had more than one allele because they consisted of multiple-linked SNPs in which the phase was determined.

Locus	No. SNPs	Type	F_{ST}	D_{EST}	H_e	No. alleles
VT	1	SNP	0.086	0.101	0.491	2
IN	2	SNP	0.053	0.015	0.228	3
SP	1	SNP	0.076	0.074	0.484	2
RH	1	SNP	0.081	0.012	0.129	2
VR	3	SNP	0.156	0.308	0.730	4
IS	2	SNP	0.081	0.120	0.622	4
ER	1	SNP	0.388	0.225	0.435	2
PL	1	SNP	0.096	0.034	0.276	2
RF	1	SNP	0.218	0.071	0.303	2
CL	1	SNP	0.034	0.020	0.408	2
PER	1	SNP	0.040	0.004	0.082	2
MT	8	SNP	0.345	0.371	0.454	15
One104	N/A	mSat	0.027	0.346	0.951	35
One102	N/A	mSat	0.011	0.127	0.922	38
Otsg68	N/A	mSat	0.019	0.298	0.956	54
SSa419	N/A	mSat	0.028	0.157	0.872	31
One114	N/A	mSat	0.017	0.178	0.933	55
Omy1011	N/A	mSat	0.026	0.272	0.937	34
One101	N/A	mSat	0.060	0.384	0.908	42
Oki100	N/A	mSat	0.044	0.313	0.905	30
Ots103	N/A	mSat	0.022	0.378	0.965	49
mSats only			0.028	0.273	0.928	
SNPs only			0.157	0.089	0.399	
Overall			0.073	0.172	0.627	

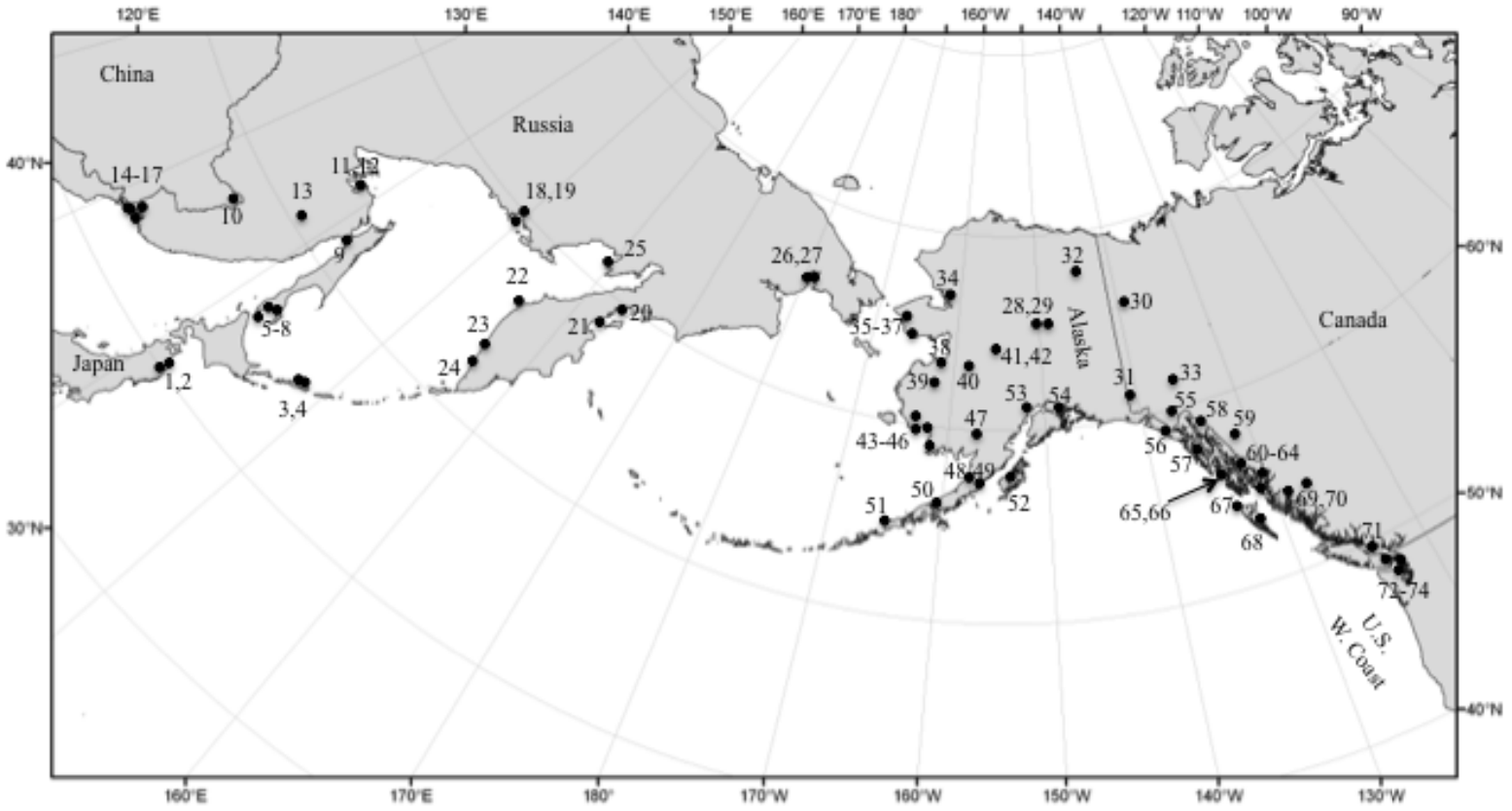


Figure 1. -- Area of study. Geographical sampling locations (dots) of the 74 stocks used for mixed-stock analysis. Regional groups are provided in Table 1.

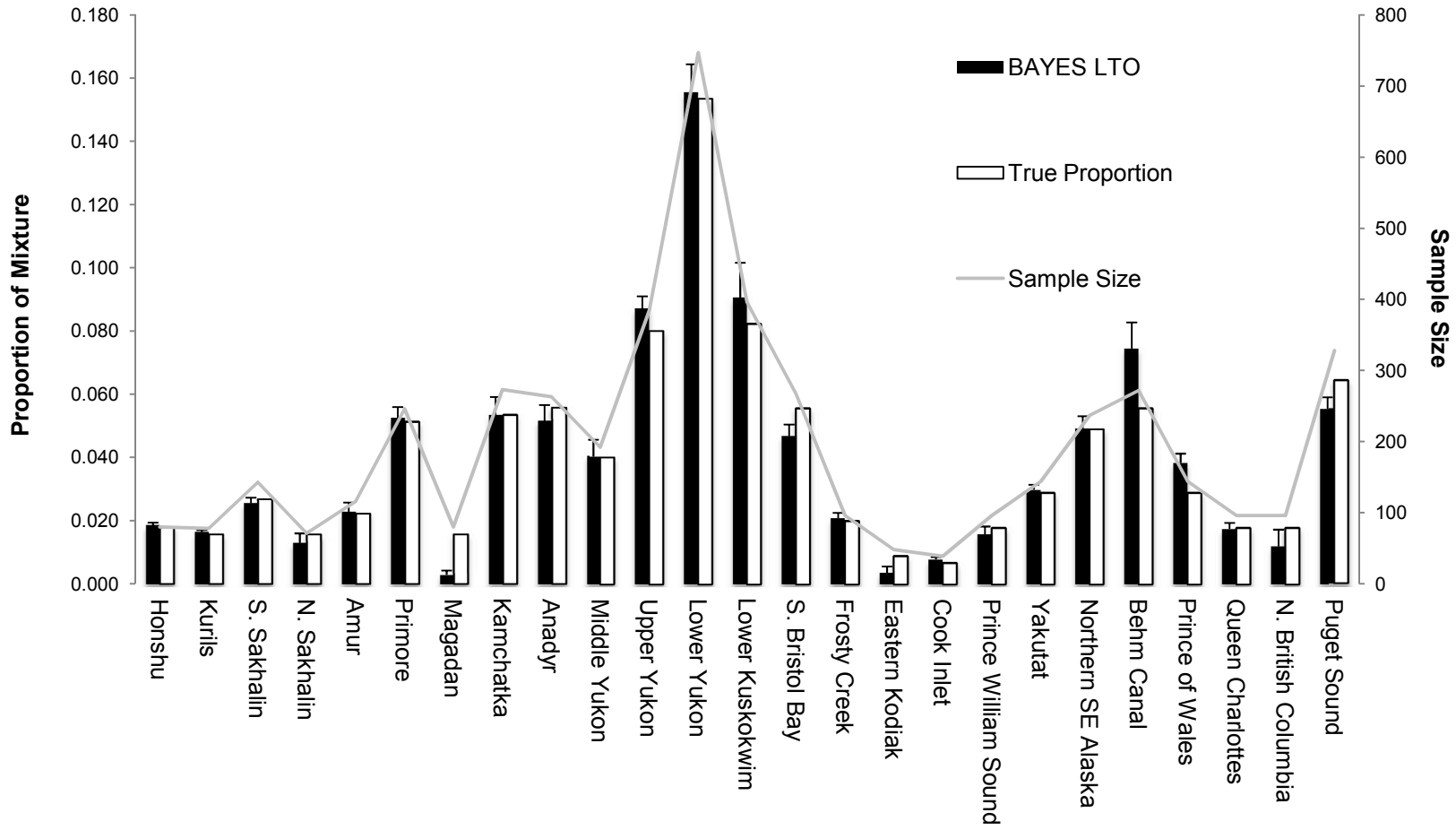


Figure 2a. -- Mean stock composition estimates for 25 reporting groups. The means (black bars) and standard errors (whiskers above the bars) were calculated from the 10 test point estimates of stock composition computed with BAYES from the combined microsatellite and SNP baseline. The unfilled bar for each region shows the true proportion of the mixtures created. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.

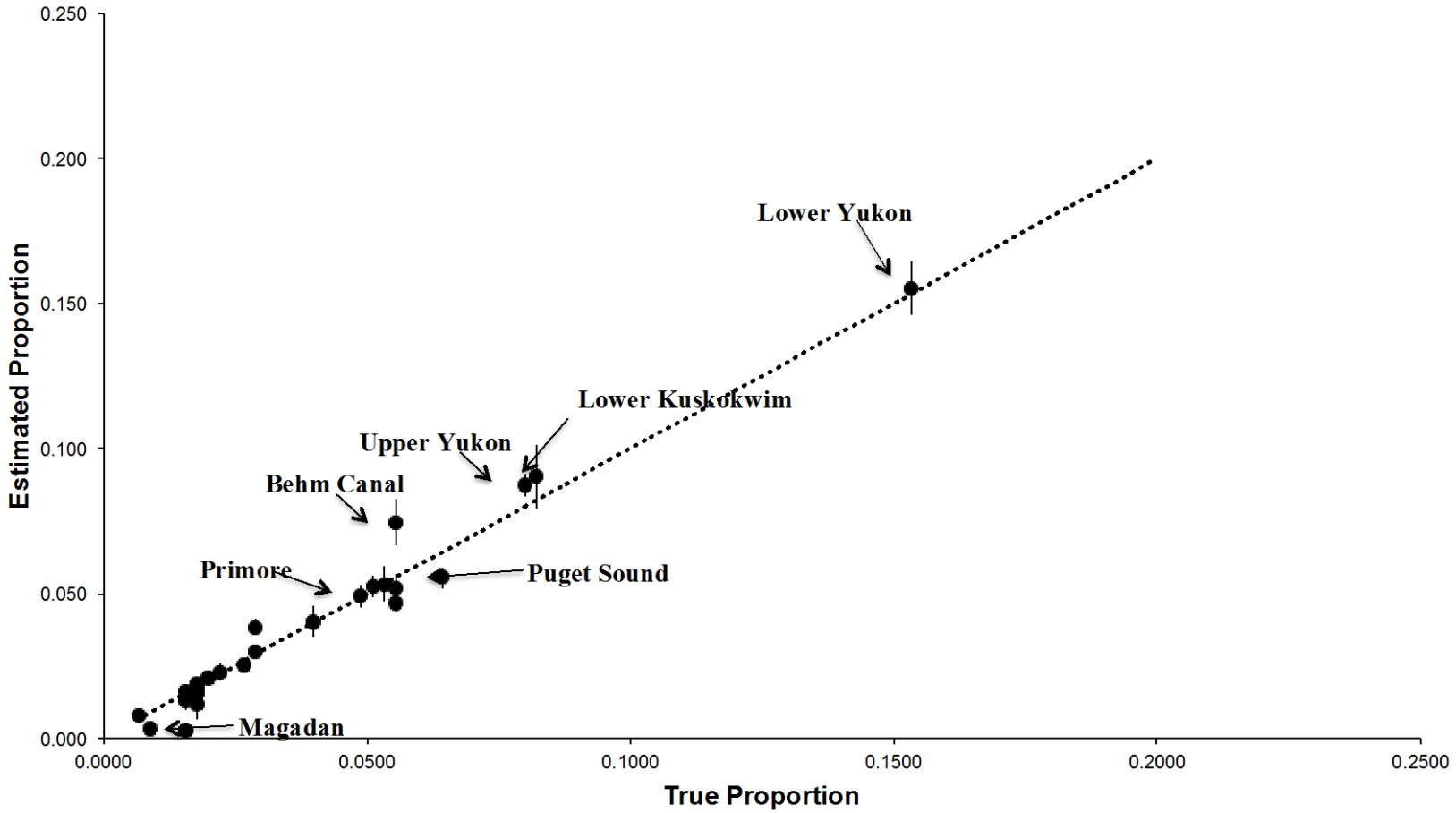


Figure 2b. -- Comparison of the mean stock composition estimates for BAYES leave-ten-percent-out cross-validation of mixtures for 25 reporting groups versus the true composition of the mixture. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value (100% correct assignment). Each circle represents the average proportion for one of the 25 reporting groups and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included a large error for at least one of the 10 mixture samples are indicated with arrows.

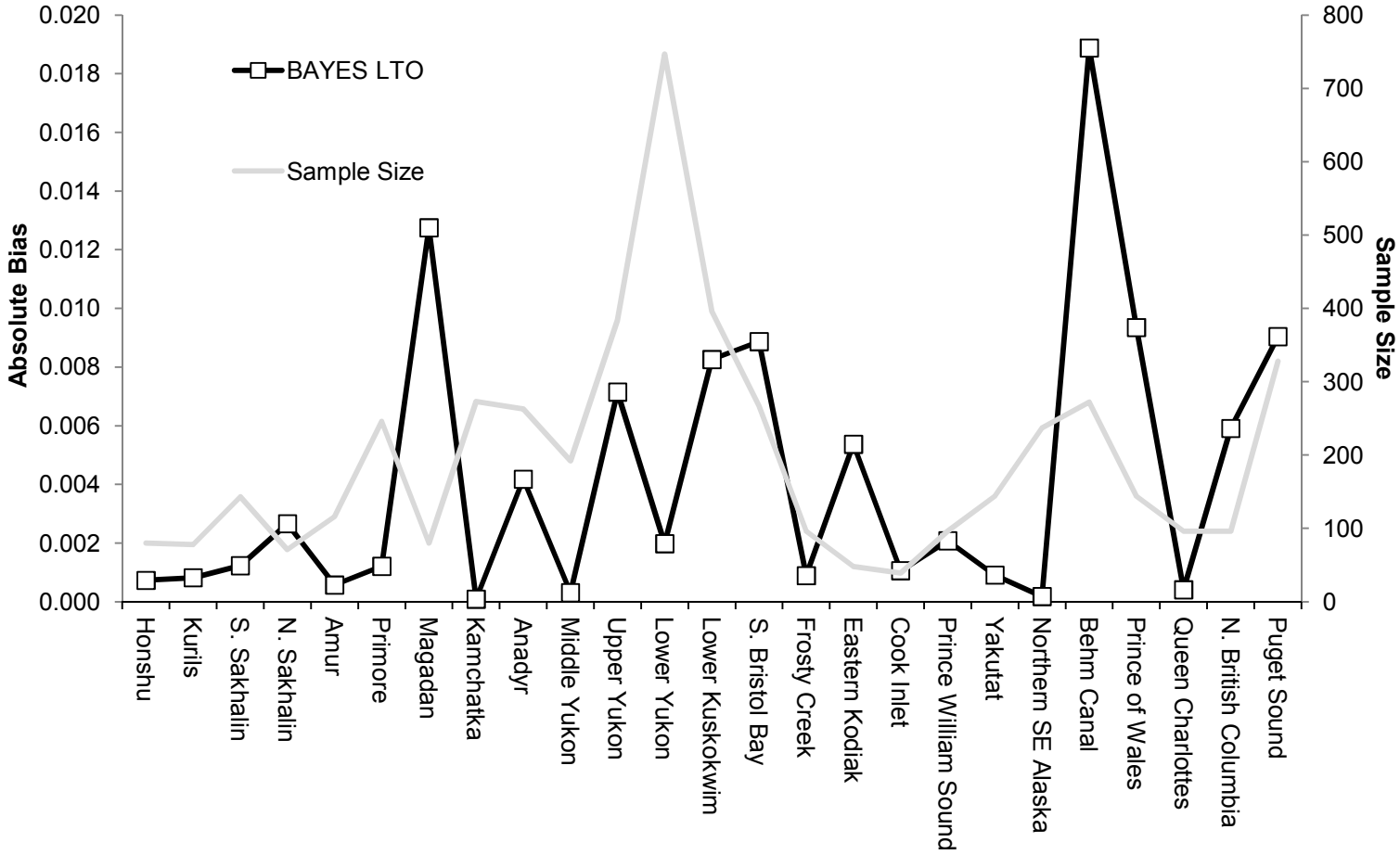


Figure 3. -- Absolute bias of composition estimates for BAYES leave-ten-percent-out cross-validation (LTO) calculated as the absolute difference between the mean estimate and the true value for each of the 25 reporting groups. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.

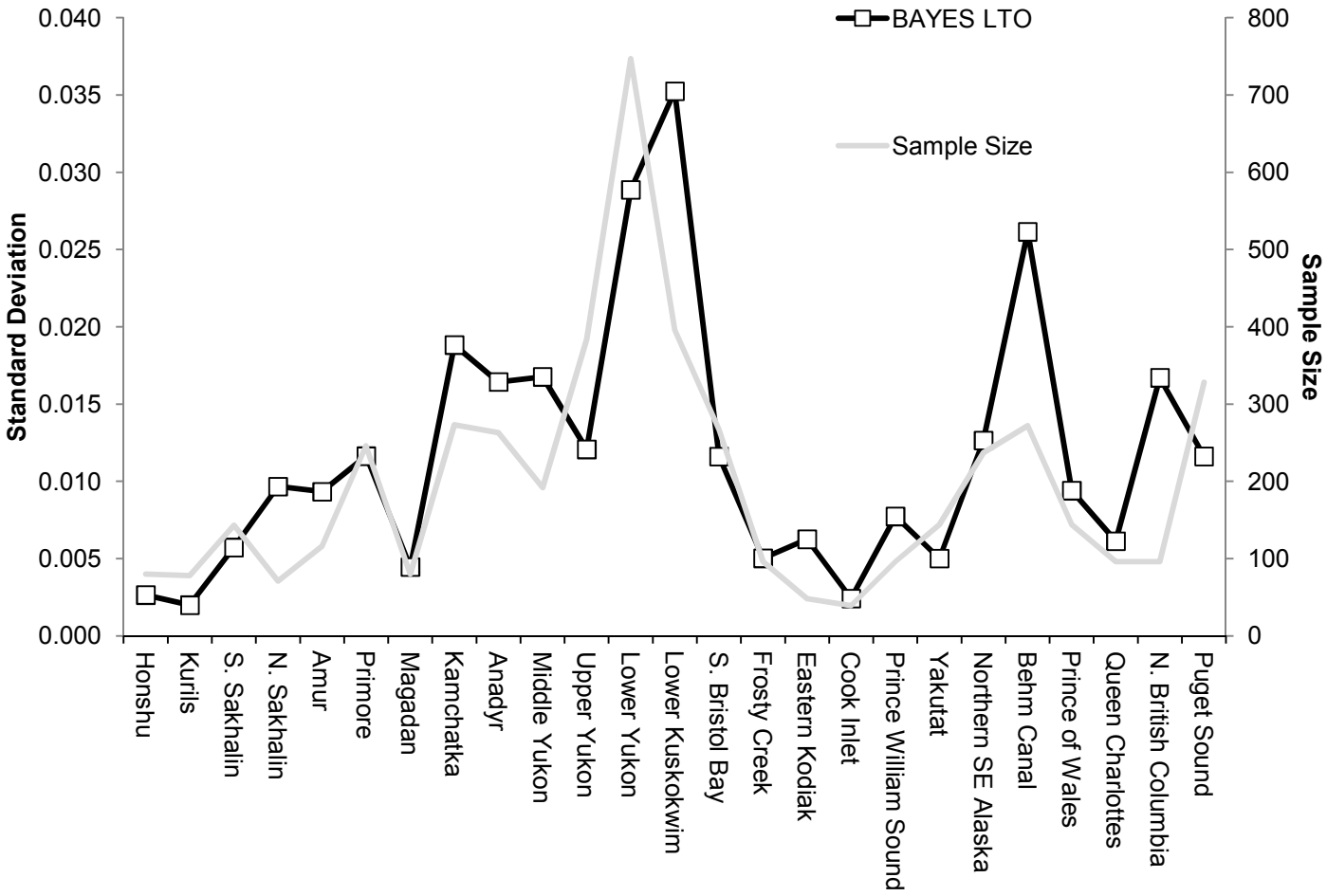


Figure 4. -- Standard deviation of stock proportion estimates for BAYES leave-ten-percent-out cross-validation (LTO). The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.

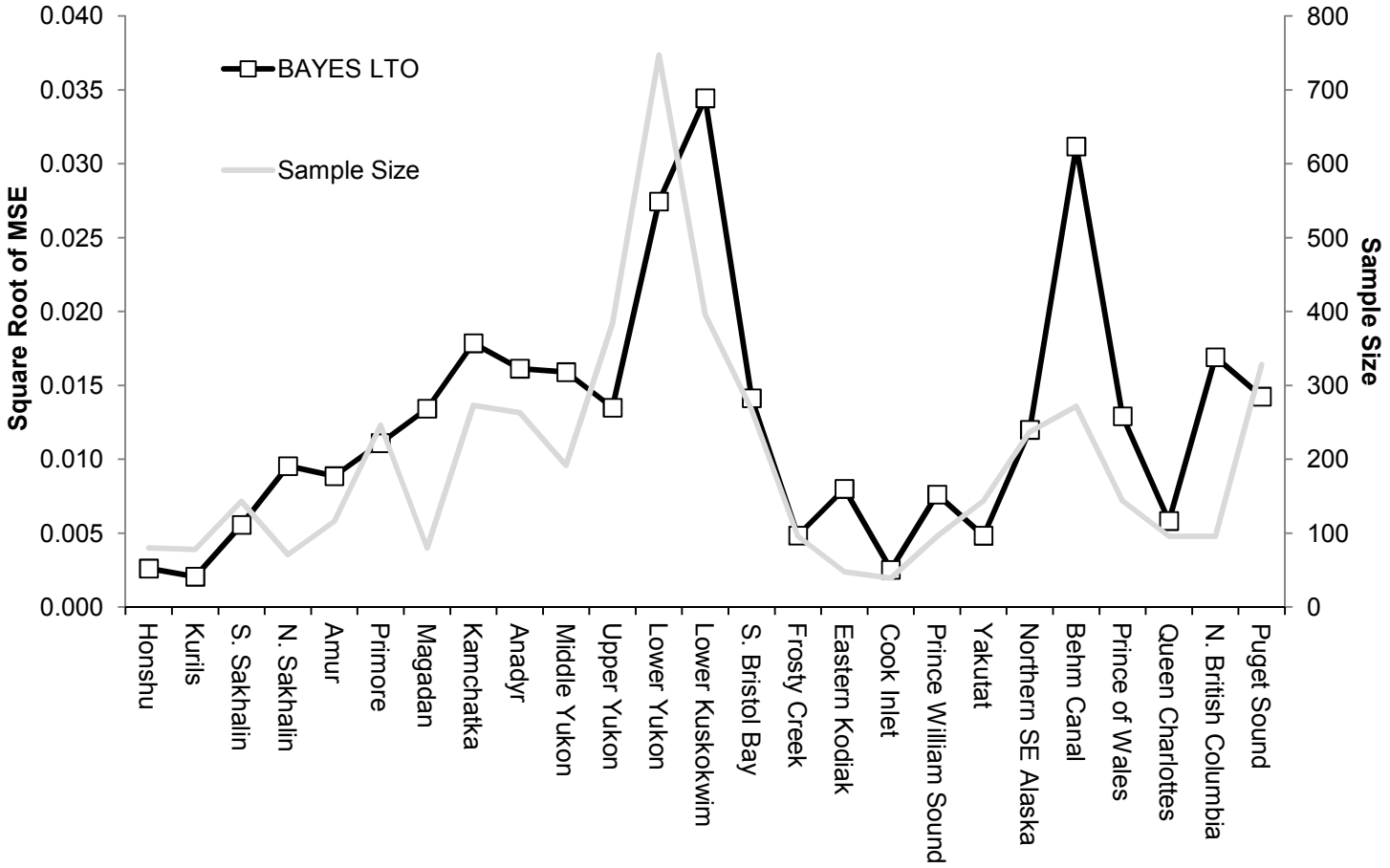


Figure 5. -- Square root of the mean-squared-error (MSE) of the stock proportion estimates for BAYES leave-ten-percent-out cross-validation (LTO). The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.

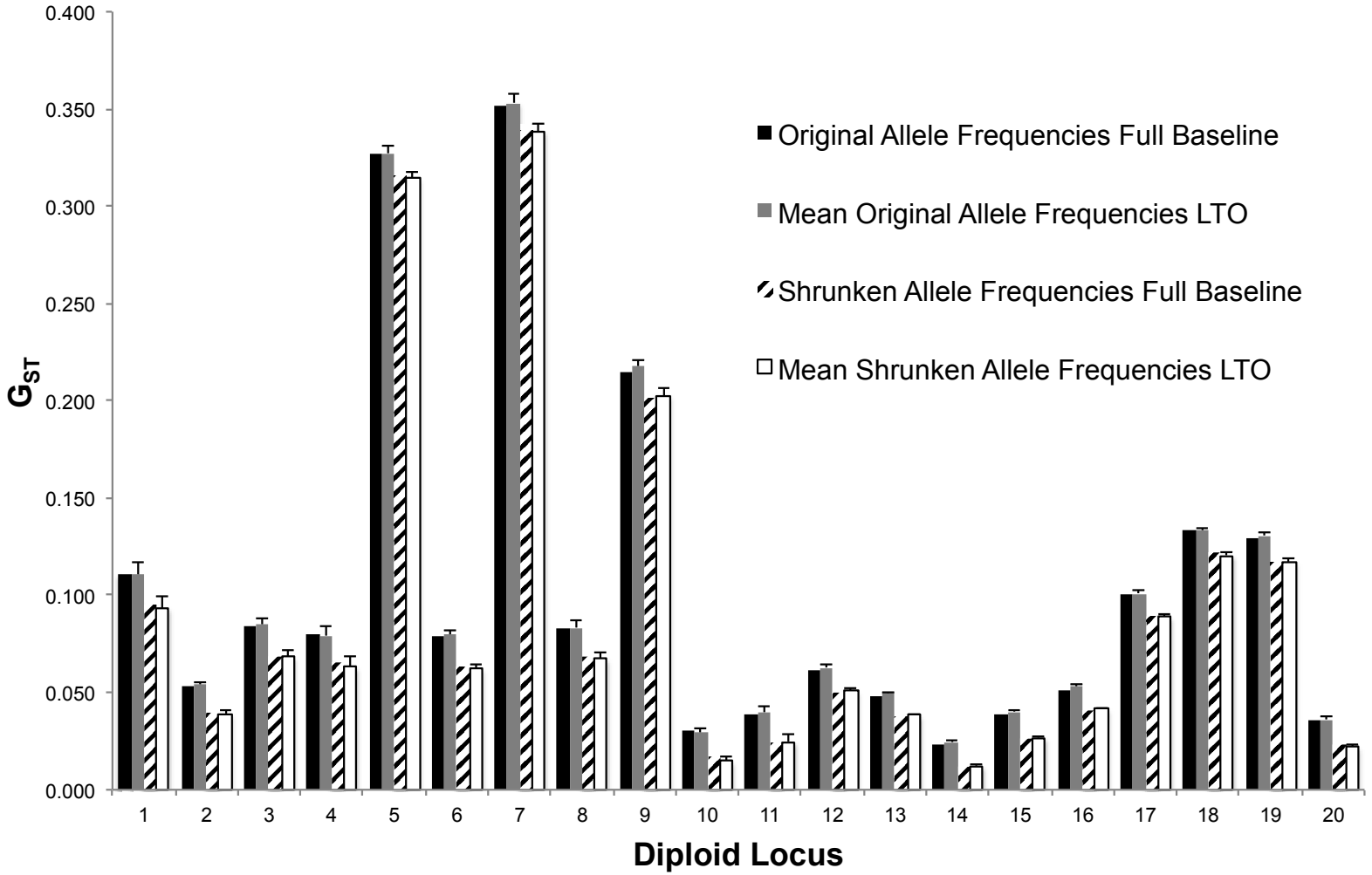


Figure 6. -- G_{ST} values to assess possible introduction of bias calculated with the observed allele frequencies from the original full baseline and then for each of the 10 leave-ten-percent-out cross-validation (LTO) baselines. The calculations were repeated with allele frequencies that were shrunk toward their prior grand mean according to Equation 4 in Pella and Masuda (2001).

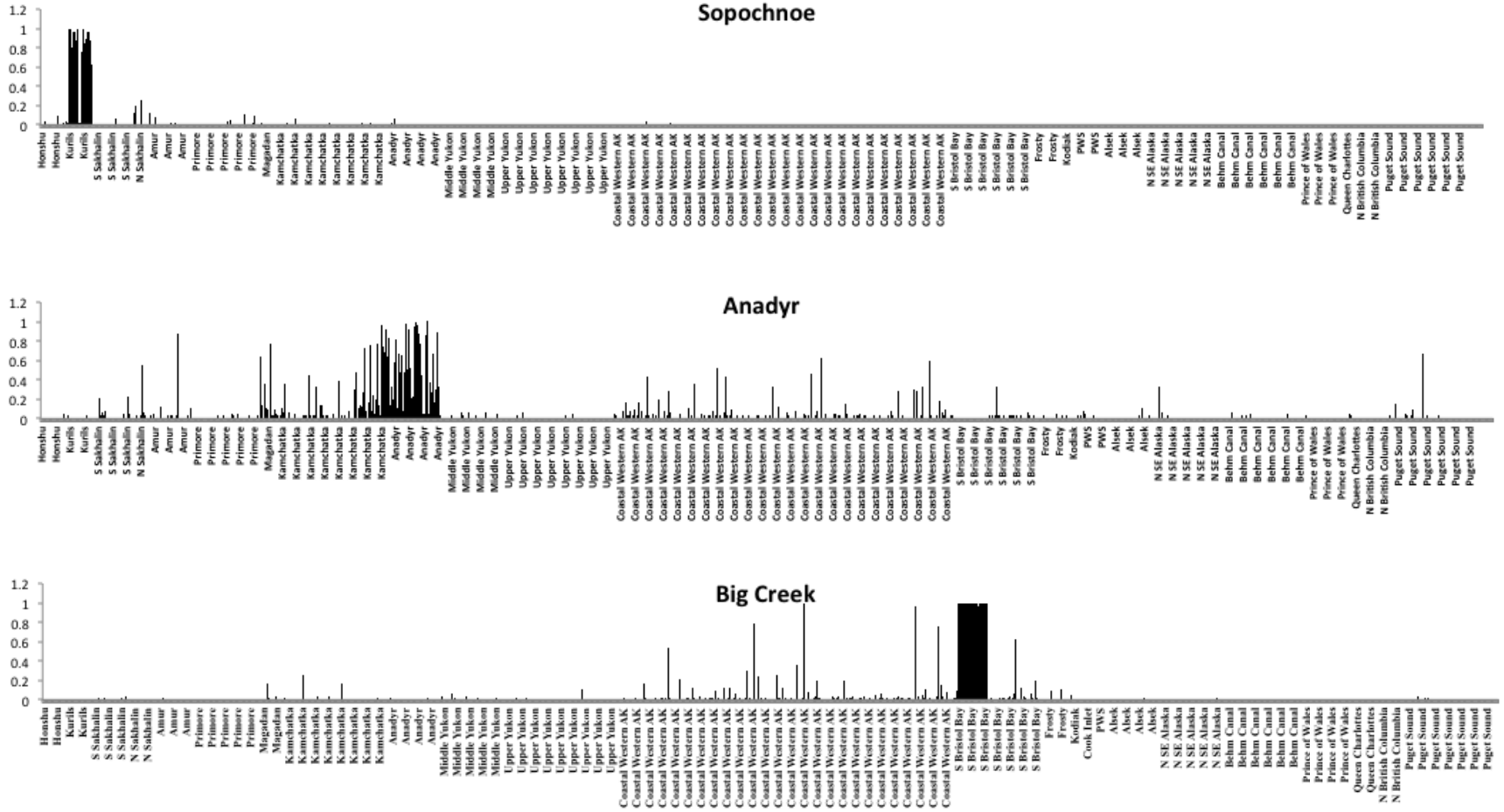


Figure 7. -- Individual assignments for 3 of the 74 populations in the baseline. Each bar represents an individual. The y-axis is the proportion of times that the individual was assigned to a population given on the x-axis. Individuals from Sopochnoe assign to the Kuril region with high probability and those in Big Creek assign to the correct drainage with high accuracy. Those in Anadyr, however, show mis-assignments to both Asian and Alaskan stocks. Data for the 74 populations are available upon request from the lead author.

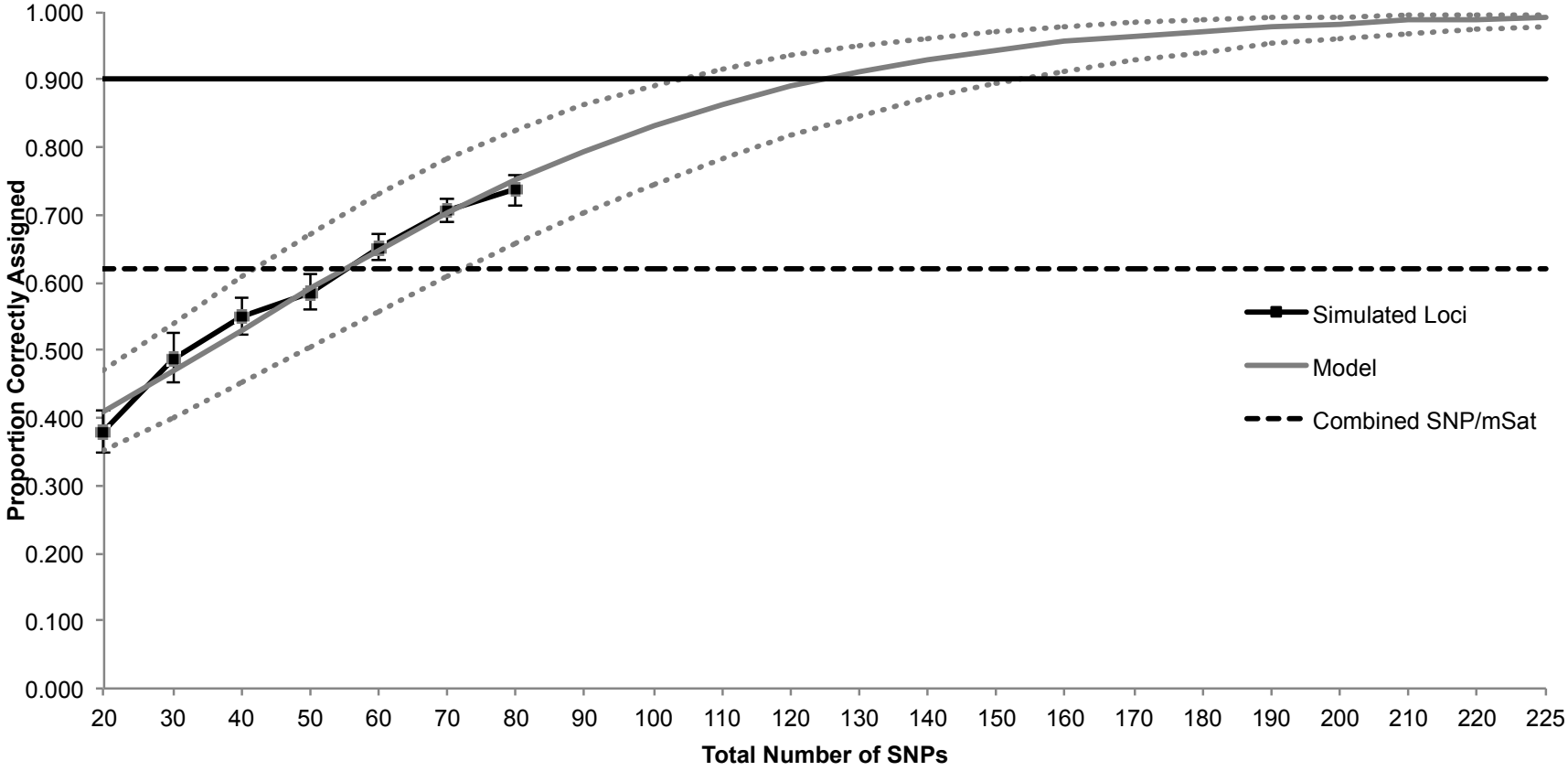


Figure 8. -- Mean proportion of individuals in the mixture correctly assigned to their reporting group with the maximum *a posteriori* rule as related to the total number of single nucleotide polymorphisms (SNPs), the original (12 SNPs), and the simulated (8 or more SNPs). Simulated data are shown with the solid line and black squares. The gray line represents a logistic function fit to the data with maximum likelihood and the dotted gray lines are the 2.5% and 97.5% confidence intervals. The solid black horizontal line represents the 0.9 proportion correctly assigned and the dashed black horizontal line is the proportion correctly assigned with the current combined SNP, microsatellite (mSat), and mitochondrial DNA data (21 loci).

APPENDIX 1

Single Nucleotide Polymorphism Discovery

The development of baselines from single nucleotide polymorphism (SNP) data involves four steps: 1) discovery of SNP loci, 2) development of a laboratory assay for each SNP, 3) selection of SNPs for inclusion into the baseline among those assayed, and 4) evaluation of the SNP baseline's capacity to distinguish its stocks in mixtures. Most baseline development projects use Sanger or next-generation sequencing to discover tens or even thousands of SNPs from which a subset is developed into a laboratory assay and used to genotype baseline samples. In an attempt to reduce the number of SNPs to be used for MSA, a subset of the most promising are high-graded and then evaluated for precision and accuracy, usually with simulations of mixture samples from the baseline itself.

Anderson (2010) cautioned that a systematic upward bias in predicted accuracy can be introduced into baseline evaluation when SNPs are high-graded if the same samples are used to choose loci and to evaluate the new baseline. This bias is distinct from error-enhanced divergence described earlier by Anderson et al. (2008) and results from the fact that divergence estimates from a sample are larger or smaller than the true value. SNP loci that are chosen based on their high divergence estimates will exhibit the statistical phenomenon called regression to the mean and likely not perform as well with different baseline and mixture samples, and loci that were initially excluded because of low divergence may perform better than expected in later samples. High-grading is also different than so-called 'ascertainment bias', which occurrs when too few individuals are used in the ascertainment panel for SNP discovery (although this is best described as sampling error, we will use the term most often reported in the literature).

We constructed an alternative system to develop and evaluate SNP baselines that reduces ascertainment bias introduced during the discovery step, reduces the costs associated with the development step as well as bias introduced during the selection step (Garvin and Gharrett 2007), and either reduces or eliminates optimistic bias in baseline evaluation from error-enhanced divergence and high-grading. Our Eco-TILLING method essentially combines the discovery and selection steps (1 and 3) of the baseline development into a single step. Our ascertainment panel consisted of 480 individuals that represented 12 populations across a geographic range, which was used to survey each target DNA sequence. Ascertainment bias was reduced because 40 individuals per stock were surveyed for genetic variants compared to a handful with standard sequencing methods. Costs were reduced because only informative SNPs are subsequently developed into laboratory assays.

We used a single population to represent each geographic reporting group for both SNP discovery and to high-grade loci with Eco-TILLING (Garvin and Gharrett 2007). Importantly, the representative population for a reporting group was not always the same for all SNPs discovered, and less than half of a sample from a population was used for the discovery and high-grading step.

For our SNP discovery efforts, we amplified targeted DNA sequences with pools of DNA from a regional representative stock and chose potentially informative loci according to estimated allele frequencies. Those loci were then evaluated with our LTO method with all of the baseline samples, which included the first portions of the sample used to choose the loci (the training set), the second portion of the sample (the holdout set), and complete samples from other populations within the geographic region that the discovery sample represented (additional holdout samples).

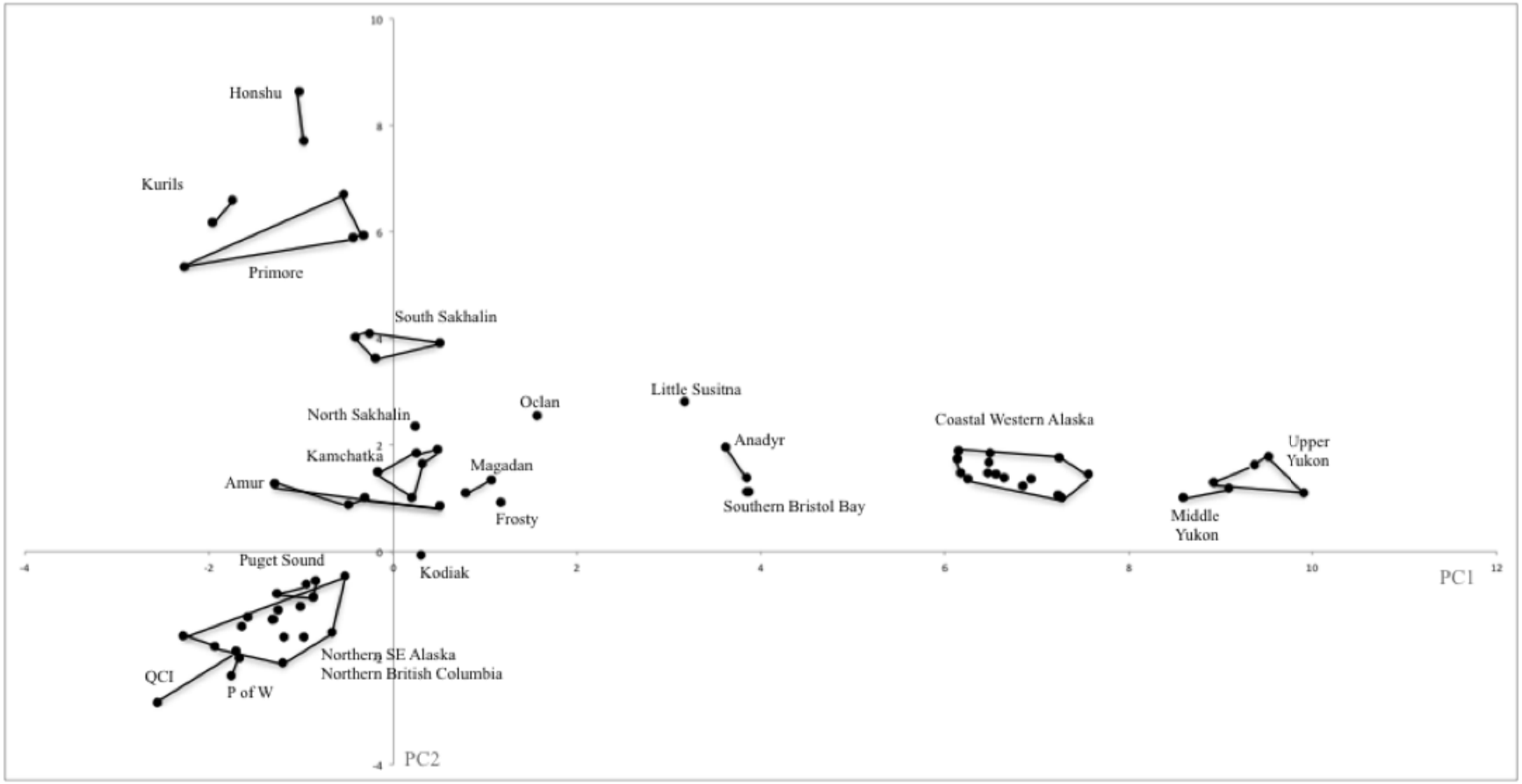
The mitochondrial SNPs were validated from previous RFLP work (Garvin et al. 2010) and other SNPs were reported in a method in which the phase of linked SNPs was determined empirically (Garvin and Gharrett 2010). Potentially informative SNPs are those with high F_{ST} values (Storer et al. 2012), or ϕ_{ST} for the mitochondrial variants (Garvin et al. 2010; $\phi_{ST} > 0.3$, Moriya et al. 2006).

Principal Components Analysis

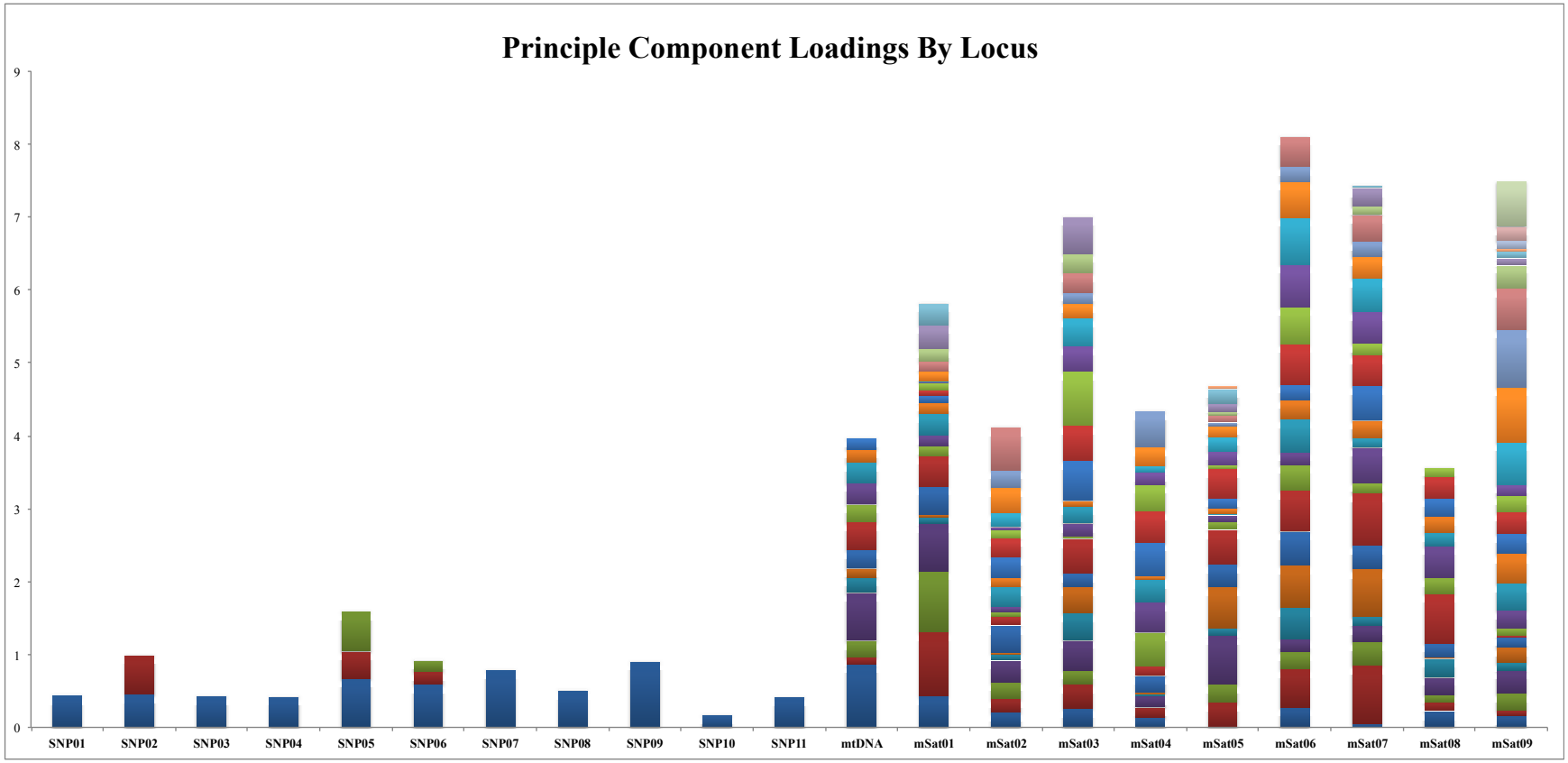
A principal components analysis (PCA) was performed on the allele frequency data that were arcsine square-root transformed before the analysis in SYSTAT (Sokal and Rohlf 1994). For the multi-allelic microsatellite data, if the expected number of alleles at each locus for the smallest sample was less than four, the alleles were binned with the next largest-sized allele. The absolute values of the loadings for the first component were also computed to quantify the informativeness of each allele at each locus.

Appendix Table 1-1. -- Mean stock proportion estimates from each of the 10 BAYES leave-ten-percent-out cross-validation analyses with the 25 reporting groups.

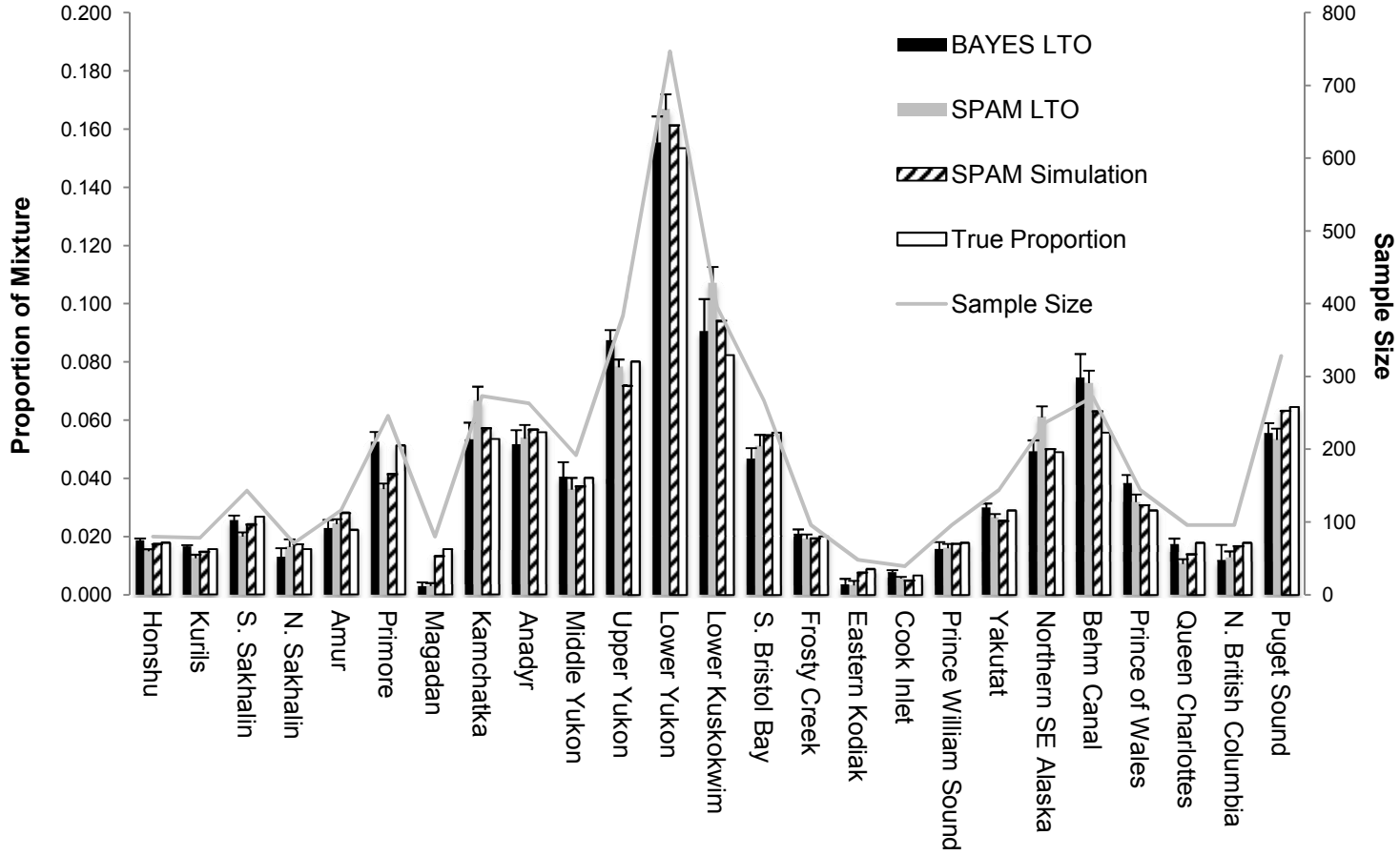
Group	Mean Run 1	SD	Mean Run 2	SD	Mean Run 3	SD	Mean Run 4	SD	Mean Run 5	SD	Mean Run 6	SD	Mean Run 7	SD	Mean Run 8	SD	Mean Run 9	SD	Mean run 10	SD	True Proportion
Honshu	0.020	0.007	0.018	0.007	0.019	0.007	0.017	0.007	0.019	0.007	0.022	0.007	0.013	0.006	0.019	0.007	0.021	0.007	0.019	0.007	**0.018**
Kurils	0.018	0.006	0.016	0.006	0.017	0.007	0.015	0.006	0.020	0.007	0.016	0.006	0.013	0.006	0.015	0.006	0.016	0.006	0.019	0.007	**0.016**
S. Sakhalin	0.036	0.011	0.028	0.009	0.030	0.010	0.022	0.008	0.024	0.009	0.030	0.010	0.023	0.008	0.020	0.008	0.016	0.007	0.028	0.011	**0.027**
N. Sakhalin	0.014	0.009	0.018	0.009	0.006	0.008	0.030	0.010	0.004	0.008	0.019	0.011	0.019	0.009	0.000	0.001	0.001	0.003	0.015	0.011	**0.016**
Amur	0.019	0.008	0.038	0.011	0.011	0.009	0.013	0.009	0.027	0.009	0.032	0.013	0.023	0.009	0.016	0.008	0.014	0.007	0.028	0.011	**0.031**
Primore	0.044	0.011	0.042	0.010	0.064	0.014	0.067	0.014	0.043	0.011	0.043	0.012	0.067	0.014	0.064	0.013	0.054	0.012	0.040	0.010	**0.042**
Magadan	0.002	0.004	0.001	0.004	0.000	0.001	0.001	0.003	0.001	0.003	0.001	0.002	0.000	0.002	0.011	0.009	0.011	0.009	0.001	0.003	**0.016**
Kamchatka	0.056	0.014	0.057	0.014	0.042	0.014	0.043	0.013	0.054	0.015	0.048	0.016	0.024	0.010	0.093	0.022	0.071	0.017	0.050	0.016	**0.053**
Anadyr	0.050	0.012	0.024	0.011	0.065	0.016	0.044	0.013	0.079	0.015	0.053	0.014	0.071	0.014	0.044	0.013	0.037	0.012	0.050	0.013	**0.056**
Middle Yukon	0.036	0.014	0.022	0.010	0.044	0.015	0.035	0.012	0.033	0.013	0.022	0.010	0.033	0.015	0.078	0.018	0.057	0.014	0.042	0.016	**0.040**
Upper Yukon	0.079	0.015	0.105	0.018	0.086	0.017	0.071	0.015	0.092	0.016	0.103	0.017	0.094	0.018	0.076	0.017	0.074	0.014	0.088	0.017	**0.080**
Lower Yukon	0.152	0.031	0.184	0.034	0.095	0.028	0.160	0.035	0.199	0.029	0.147	0.028	0.178	0.034	0.139	0.030	0.135	0.031	0.162	0.033	**0.153**
Lower Kuskokwim	0.112	0.031	0.083	0.032	0.154	0.031	0.110	0.035	0.019	0.019	0.100	0.027	0.066	0.033	0.064	0.027	0.107	0.031	0.095	0.035	**0.082**
S. Bristol Bay	0.047	0.013	0.039	0.012	0.034	0.013	0.055	0.015	0.062	0.014	0.045	0.012	0.049	0.014	0.049	0.014	0.064	0.015	0.033	0.015	**0.056**
Frosty	0.020	0.008	0.019	0.007	0.026	0.008	0.023	0.008	0.025	0.008	0.021	0.008	0.028	0.009	0.012	0.007	0.015	0.007	0.022	0.008	**0.020**
Kodiak	0.011	0.011	0.001	0.003	0.003	0.004	0.002	0.004	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.002	0.016	0.010	**0.009**
Cook Inlet	0.007	0.004	0.007	0.004	0.010	0.006	0.003	0.004	0.009	0.005	0.009	0.005	0.012	0.006	0.008	0.005	0.006	0.004	0.006	0.004	**0.007**
Prince William Sound	0.020	0.011	0.017	0.008	0.001	0.004	0.019	0.010	0.009	0.007	0.019	0.009	0.022	0.008	0.005	0.007	0.024	0.010	0.020	0.010	**0.018**
Yakutat	0.032	0.009	0.027	0.009	0.021	0.007	0.024	0.008	0.035	0.010	0.031	0.009	0.036	0.009	0.028	0.009	0.031	0.009	0.034	0.009	**0.029**
Northern SE AK	0.055	0.015	0.059	0.016	0.059	0.015	0.047	0.016	0.059	0.016	0.054	0.015	0.036	0.013	0.031	0.013	0.061	0.017	0.038	0.019	**0.049**
Behm Canal	0.039	0.021	0.103	0.030	0.090	0.021	0.062	0.018	0.054	0.015	0.060	0.020	0.057	0.016	0.124	0.024	0.083	0.021	0.060	0.026	**0.056**
Prince of Wales	0.045	0.014	0.027	0.012	0.033	0.013	0.047	0.015	0.031	0.017	0.033	0.017	0.057	0.015	0.029	0.015	0.036	0.016	0.043	0.016	**0.029**
Haida Gwaii	0.013	0.007	0.016	0.011	0.018	0.008	0.029	0.011	0.014	0.008	0.008	0.011	0.023	0.010	0.014	0.008	0.017	0.007	0.020	0.011	**0.018**
Northern B.C.	0.020	0.015	0.008	0.014	0.004	0.007	0.003	0.006	0.032	0.012	0.050	0.015	0.000	0.001	0.002	0.005	0.002	0.006	0.008	0.013	**0.018**
Puget Sound	0.053	0.014	0.043	0.015	0.069	0.014	0.057	0.014	0.058	0.012	0.036	0.014	0.057	0.013	0.061	0.015	0.048	0.013	0.067	0.017	**0.064**



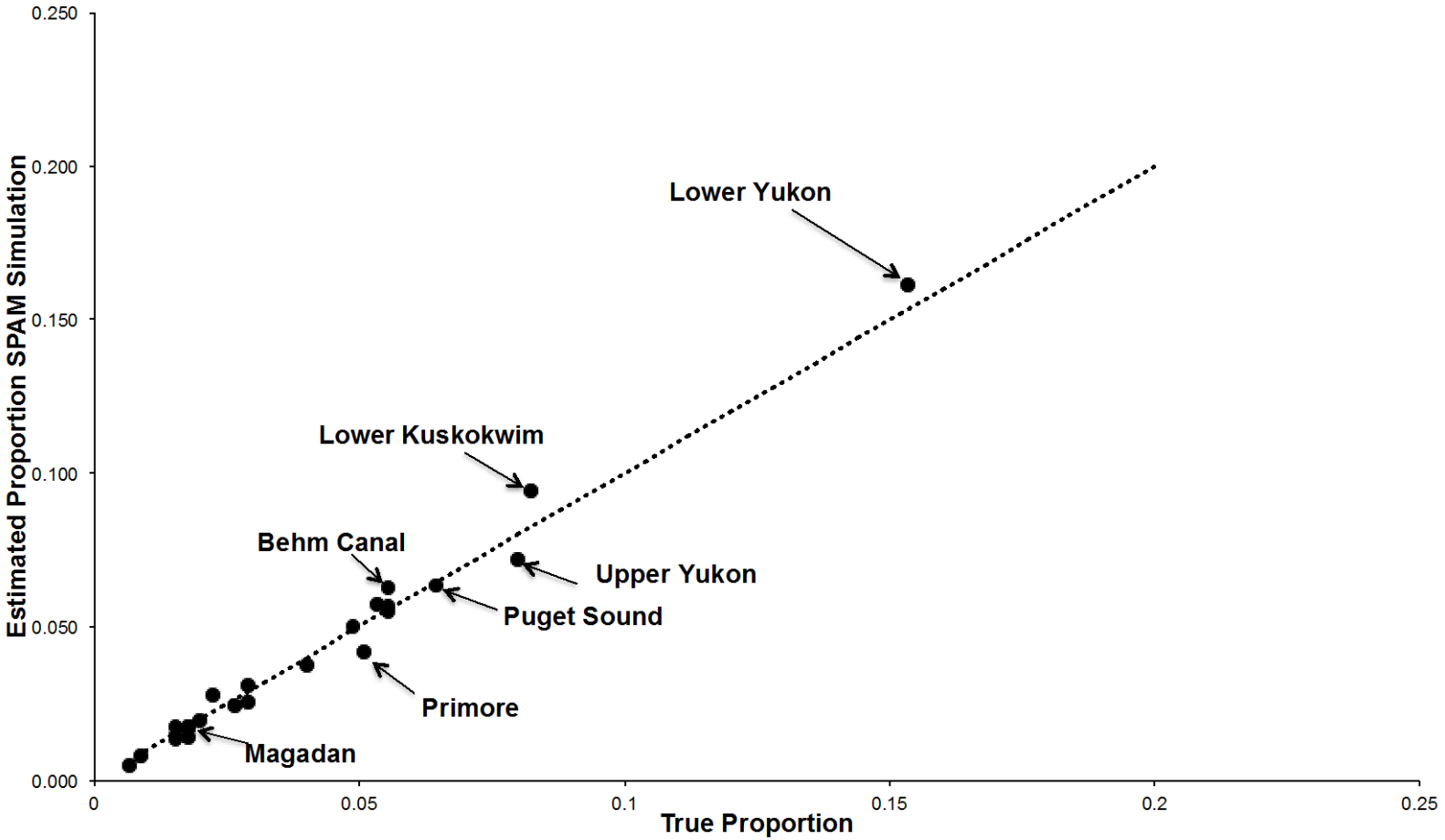
Appendix Figure 1-1a. -- Principal components analysis of the 74 baseline stocks. The first and second components are shown and lines indicate geographical groups.



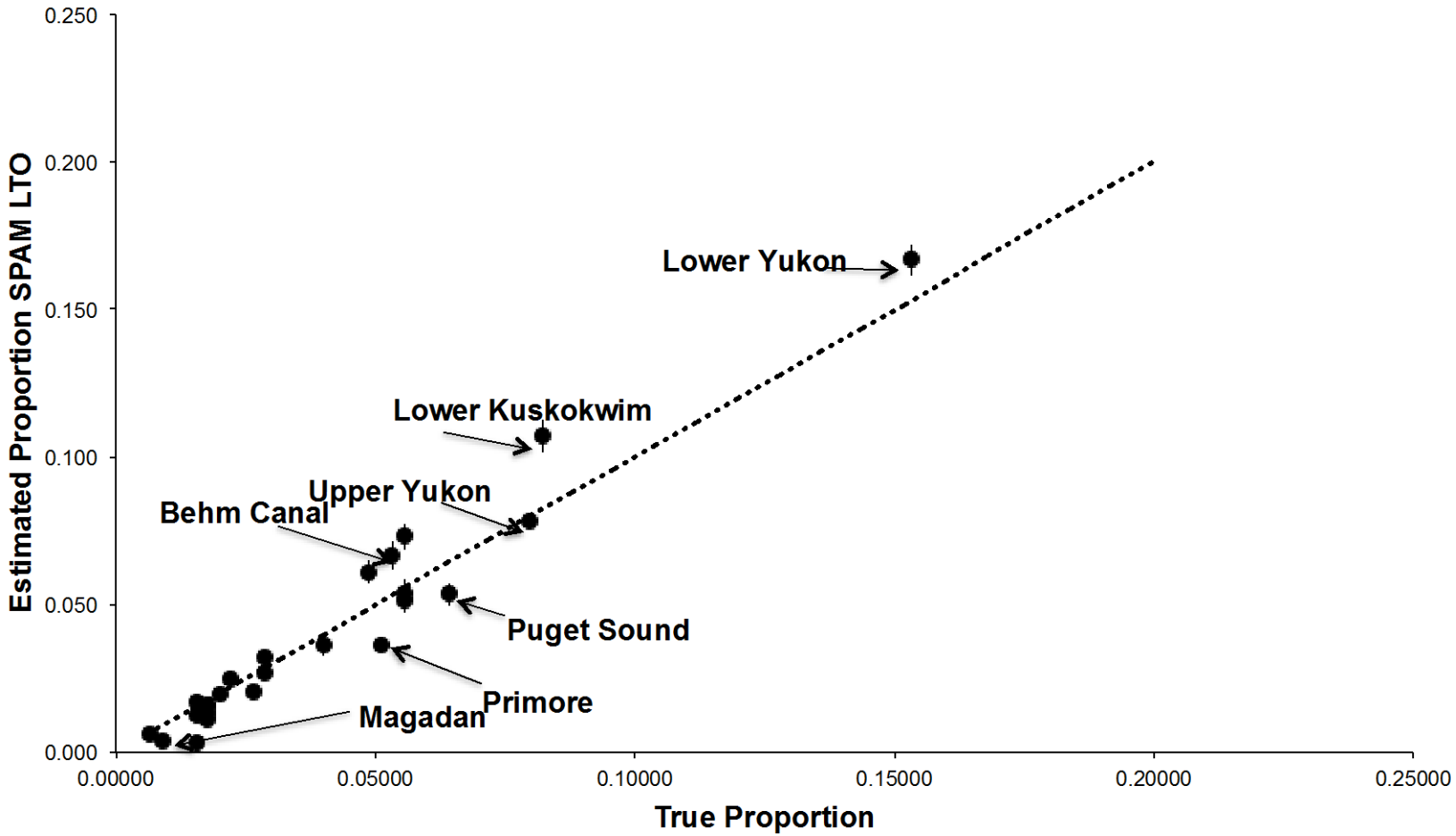
Appendix Figure 1-1b. -- The loadings for each of the 21 loci used in this study. Loadings for each allele are identified in each locus with color-scale bars.



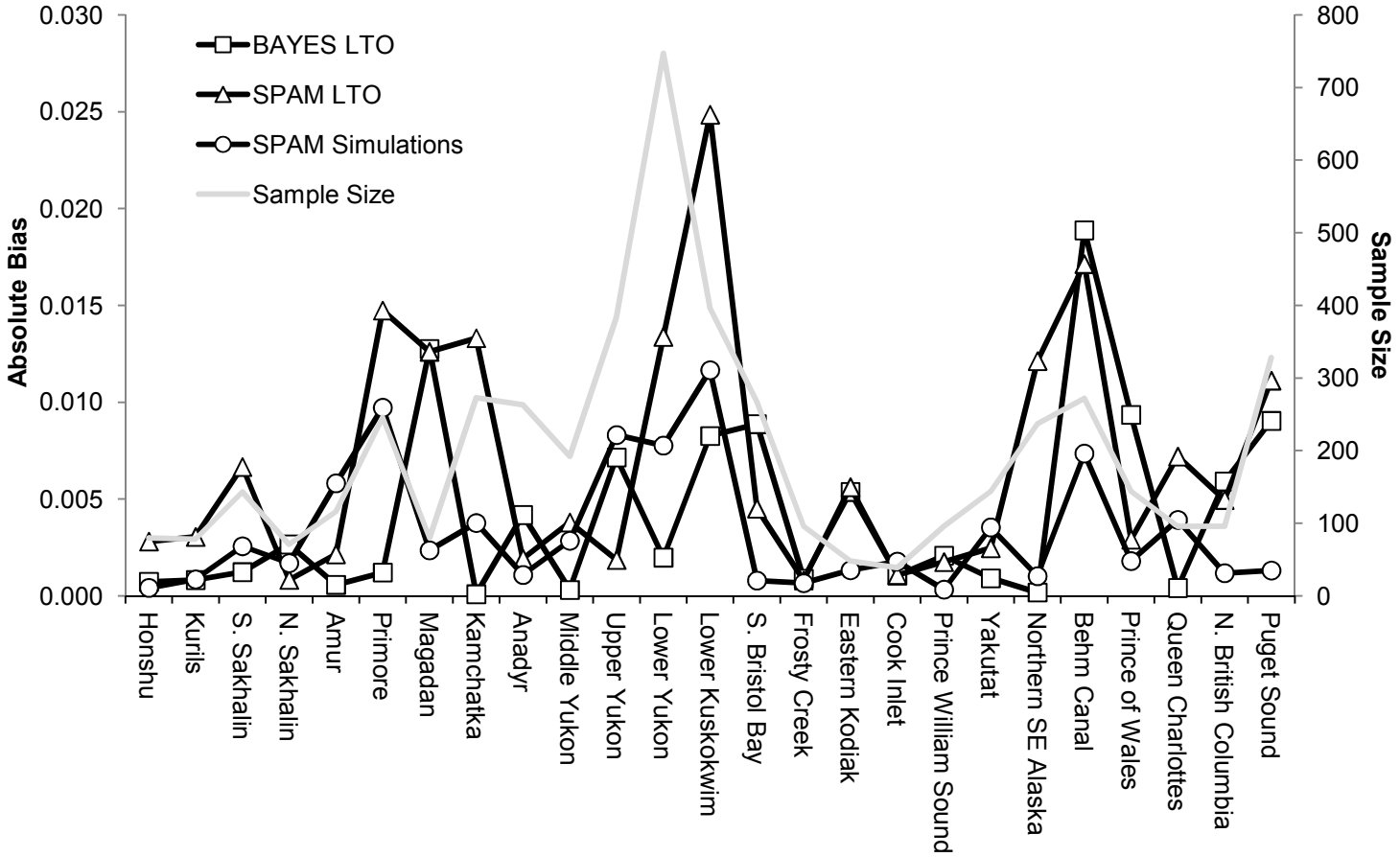
Appendix Figure 1-2. -- Mean stock composition estimates for 25 reporting groups with all three methods: BAYES leave-ten-percent-out cross-validation (LTO), SPAM LTO, and SPAM Simulation. The means (bars) and standard errors (whiskers above the bars) were calculated from the 10 test point estimates of stock composition computed with BAYES from the combined microsatellite and single nucleotide polymorphism baseline. The unfilled bar for each region shows the true proportion of the mixtures created. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.



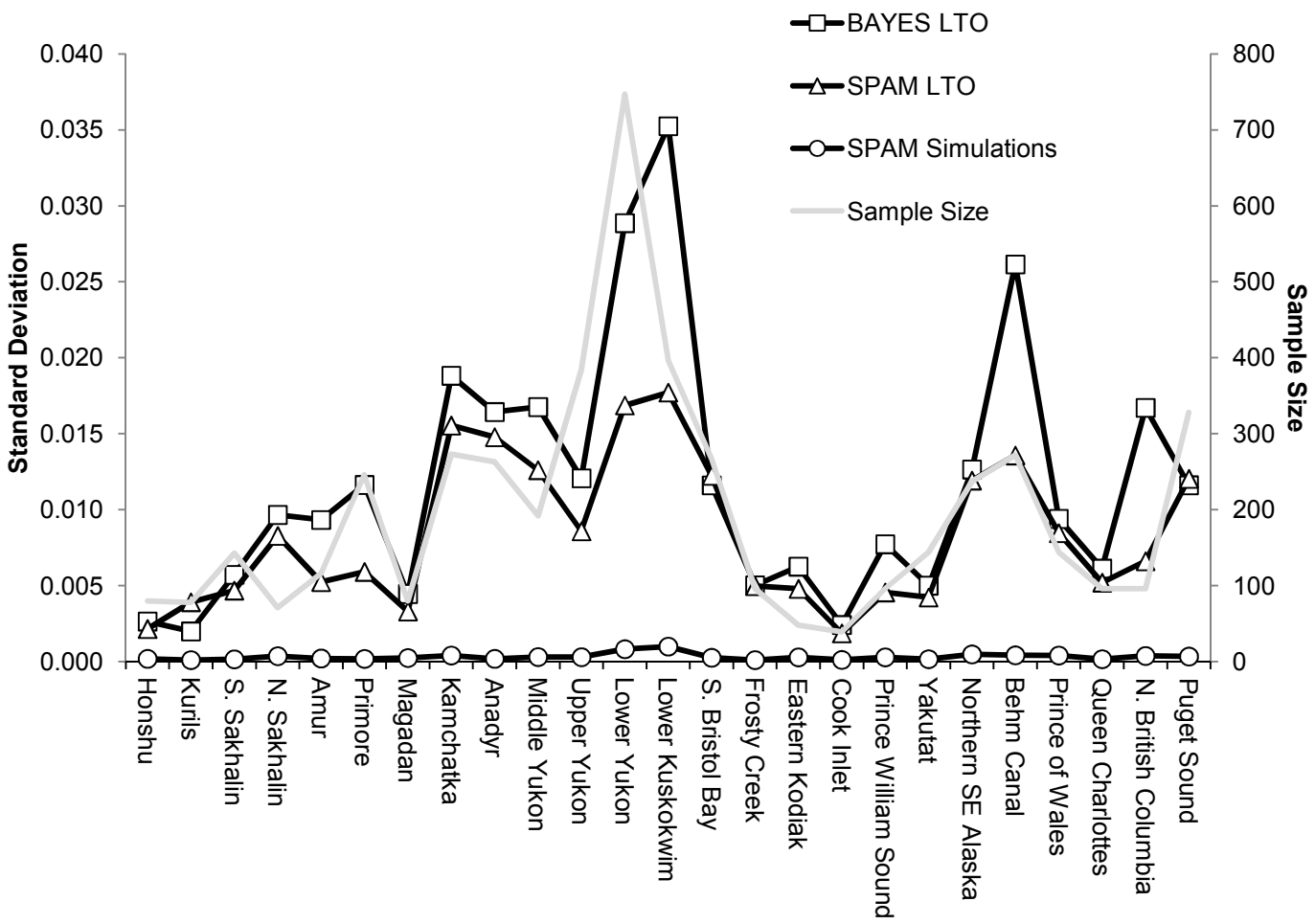
Appendix Figure 1-3. -- Comparison of the mean stock composition estimates for SPAM Simulation of mixtures for 25 reporting groups versus the true composition of the mixture. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value (100% correct assignment). Each circle represents the average proportion for one of the 25 reporting groups, and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included a large error for at least one of the ten mixture samples are indicated with arrows.



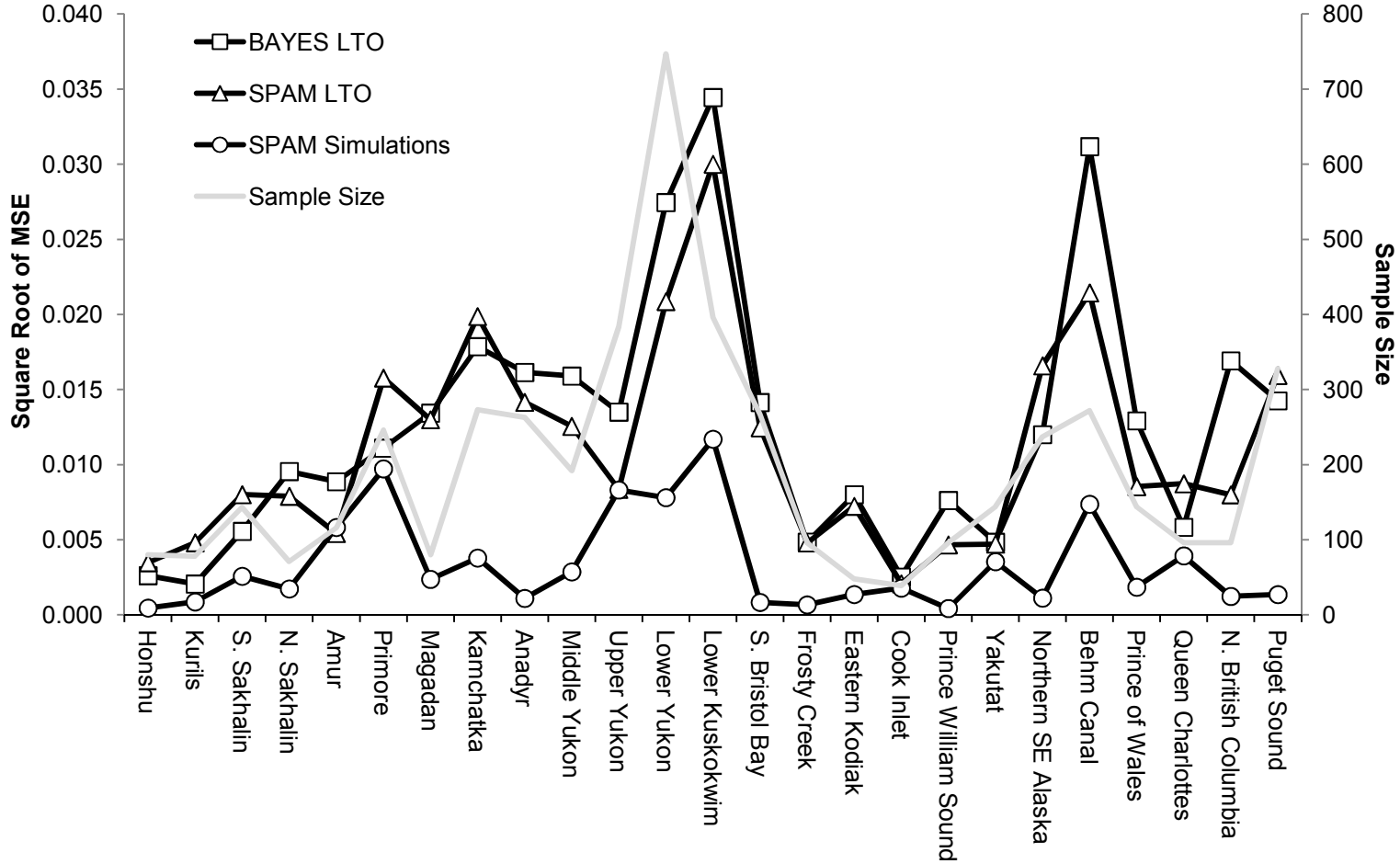
Appendix Figure 1-4. -- Comparison of the mean stock composition estimates for SPAM leave-ten-percent-out cross-validation (LTO) of mixtures for 25 reporting groups versus the true composition of the mixture. The black diagonal line represents the relationship between a perfectly accurate estimate and the true value (100% correct assignment). Each circle represents the average proportion for one of the 25 reporting groups, and the standard error of the estimated proportions is indicated by the whiskers for each circle. Names of groups whose averages included a large error for at least one of the ten mixture samples are indicated with arrows.



Appendix Figure 1-5. -- Absolute bias of composition estimates for BAYES leave-ten-percent-out cross-validation (LTO), SPAM LTO, and SPAM Simulation calculated as the absolute difference between the mean estimate and the true value for each of the 25 reporting groups. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.



Appendix Figure 1-6. -- Standard deviation of stock proportion estimates for BAYES leave-ten-percent-out cross-validation (LTO), SPAM LTO, and SPAM Simulation. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.



Appendix Figure 1-7. -- Square root of the mean-squared-error (MSE) of the stock proportion estimates for BAYES leave-ten-percent-out cross-validation (LTO), SPAM LTO, and SPAM Simulation. The gray continuous line is scaled by the secondary y-axis and shows the sample size for each group.

APPENDIX 2

```
### This code describes the leave-ten-percent-out cross-validation for mixed-stock analysis.
### The code was written by Garvin for his Ph.D. thesis and was substantially
### altered by Patrick Barry as a side project.
### The code was written for PCs but should be able to run on
### both linux and mac OS as there are not any packages or system commands
### implemented.
### There are three major pieces to the code:
### Piece one makes many baseline.bse files
### Piece two makes many mixture files
### Piece three makes a control file to run with the program BAYES
### Because Bayes is extremely menu driven it is not possible to run
### it in batches from MSDOS command prompt.

### In order to execute the code you will need to have your data
### formatted according to the example given:
### The first column is the Individual #
### The second column is population of the individual
### Then each column after is a single allele from a locus
### diploid loci will occupy two columns and should come before haploid loci
### Haploid loci should occupy two columns with the second column
### coded with 888 denoting that it is haploid.


### To make the control file, a Stock.ID tab delim .txt file is needed
### to specify the priors for the dirichlet and other stock
### specific options for the program Bayes.

BaseMix_csv<-function(x,Prop=10,loci.diploid,loci.haploid,print_pop.sum = TRUE)
{
 #This is a function to produce baseline and mixture files
 #from an input file. These files can then be formatted
 #for analysis with Bayes

 #read in data file
 Y<-read.csv(x,na.string="?")

 #Recode all the 888 for haploid loci with na
 Y[Y==888]<-NA

 #How many populations do we have?
 Pops <- length(unique(Y$Population))
 #How many individuals per pop?
 Pop.sum<-matrix(data=NA,ncol=4,nrow=Pops)
```

```
  Pop.sum[,1]<-seq(from=1, to = Pops,by=1)
  for (q in 1:Pops){
    Pop.sum[q,2]<-nrow(Y[Y$Population==q,,])
  }

  #We can add to our vector how many individuals to be taken that are 1/10th the population
  #and the leftovers to be added back
  for(w in 1:Pops){
    Pop.sum[w,3]<-Pop.sum[w,2]%/%10
    Pop.sum[w,4]<- Pop.sum[w,2]%%10
  }
  colnames(Pop.sum)<-c("Pop","n","N","LO")

  if (print_pop.sum == TRUE) {write.table(Pop.sum,quote=FALSE,file="Pop.sum.txt")
                              cat("Population Summary Printed to Pop.sum\n", sep="")
  }

  ###How many unique alleles for each locus?
  Allele.sum<- matrix(data=NA,nrow=loci.diploid+loci.haploid,ncol=2) #matrix to hold
information to reference later
  colnames(Allele.sum)<- c("Locus", "Max.Num")

  for (a.s in 1:nrow(Allele.sum)){ # fill in loci names
    Allele.sum[a.s,1]<- colnames(Y[a.s*2+1])
    all.alleles<-
list((as.vector(na.omit(suppressWarnings(as.numeric(rownames(table(Y[,(a.s*2+1):(a.s*2+2)])))
))))),(as.vector(na.omit(suppressWarnings(as.numeric(colnames(table(Y[,(a.s*2+1):(a.s*2+2)]))))
))))
    length(all.alleles)

    u.all<-(unique(Y[,a.s*2+1]))
    u.all<-append(u.all,unique(Y[,a.s*2+2]))
    u.all<-unique(u.all)
    u.all<-sort(u.all)
    Allele.sum[a.s,2] <- length(u.all)
    assign(paste("Locus",a.s,sep=""),u.all)# make a vector for each with a list of the alleles found
  }

  write.table(Allele.sum,file="LociSummary.csv",quote=F,sep=",", row.names=F,col.names=F)

### Create indexing in the Y file
  Y$Index<-NA # Create index vector
  index<- vector()
  for (v in 1:Pops){
    index[(sum(Pop.sum[Pop.sum[,1] < v,2])+1):(as.numeric((sum(Pop.sum[Pop.sum[,1] <
v,2])+Pop.sum[v,3]*10)))]<- rep(1:10,rep((Pop.sum[v,3]),10))
```

```
    if (as.logical(Pop.sum[v,4]>0)){
      index[(as.numeric((sum(Pop.sum[Pop.sum[,1] <
v,2])+Pop.sum[v,3]*10))+1):(as.numeric((sum(Pop.sum[Pop.sum[,1] <
v,2])+Pop.sum[v,3]*10)+Pop.sum[v,4]))]<- rep(11,Pop.sum[v,4])
    }
  }
  Y$Index<- index


  #add Index shuffle values to matrix
  Y$Index.shuff<-NA
  Index.shuff.temp<- vector(mode="numeric", length=sum(Pop.sum[,2]))
  #Shuffle the index within each population
  for (u in 1:Pops){
    P.temp<- Y[Y$Population==u,]### This gives me the data I want to use
    Index.shuff<-P.temp$Index
    Index.shuff<- sample(Index.shuff,length(Index.shuff),replace=FALSE)
    #Write the shuffled values back to Y
    Index.shuff.temp[(sum(Pop.sum[Pop.sum[,1] <
u,2])+1):(as.numeric((sum(Pop.sum[Pop.sum[,1] < u,2])+Pop.sum[u,2])))]<- Index.shuff
  }
  Y$Index.shuff<- Index.shuff.temp


  #Recombine all 101 x 10 mixtures to create 10 mixed stocks:
  for (i in 1:10){
    assign(paste('Mix',i,sep=""),Y[Y$Index==i,])
    assign(paste('Baseline',i,sep=""),Y[Y$Index!=i,])
  }

  #Clean up file by getting rid of shuffling col.
  for (i in 1:10){

assign(paste('Baseline',i,sep=""),eval(parse(text=paste('Baseline',i,'[,1:(ncol(',(paste('Baseline',i,s
ep="")),')-2)]',sep=""))))
  }

  #Export the files:
  for (i in 1:10){

write.table(eval(parse(text=paste('Mix',i,sep=""))),quote=FALSE,file=paste('Mix',i,'.csv',sep=""),
row.names = FALSE, sep=",")
  }

  for (i in 1:10){
```

```
write.table(eval(parse(text=paste("Baseline",i,sep=""))),quote=FALSE,file=paste("Baseline",i,".c
sv",sep=""),row.names = FALSE, sep=",")
  }
  cat('Baseline and Mixture files written to working directory')

}

##############################################################################
#RtoBAYESBASE
Bayes_format<-function(x,cores=2,loci.haploid,loci.diploid,npops){
  # x is the original baseline to test
  # cores = number of cores on the computer
  # L is number of loci
  # requires that .csv files produced by BaseMix_csv()
  # be located in the working directory

  #RtoBAYESBASE
  ####Read baseline file
  library(doSNOW)
  library(foreach)
  cl<-makeCluster(cores)
  registerDoSNOW(cl)

  Y<-read.csv(x,na.string="?")
  L<-loci.haploid + loci.diploid
  Loci.names<-paste("Locus",seq(from=1,to=L,by=1),sep="")


  ###How many unique alleles for each locus?
  Allele.sum<- matrix(data=NA,nrow=L,ncol=2) #matrix to hold information to reference later
  colnames(Allele.sum)<- c("Locus", "Max.Num")

  for (a.s in 1:nrow(Allele.sum)){ # fill in loci names
    Allele.sum[a.s,1]<- colnames(Y[a.s*2+1])
    all.alleles<-
list((as.vector(na.omit(suppressWarnings(as.numeric(rownames(table(Y[,(a.s*2+1):(a.s*2+2)])))
)))),(as.vector(na.omit(suppressWarnings(as.numeric(colnames(table(Y[,(a.s*2+1):(a.s*2+2)])))))
))))
    length(all.alleles)

    u.all<-(unique(Y[,a.s*2+1]))
    u.all<-append(u.all,unique(Y[,a.s*2+2]))
    u.all<-unique(u.all)
    u.all<-sort(u.all)
    Allele.sum[a.s,2] <- length(u.all)
```

```
  assign(paste("Locus",a.s,sep=""),u.all)# make a vector for each with a list of the alleles found
 }

 cat('Please be patient, this may take a moment\n')

 foreach (o = 1:10,.export=Loci.names) %dopar% {
  Base.temp<-read.csv(paste("Baseline",o,".csv", sep=""))
  #Calculate the number of populations
  P<-unique(Base.temp$Population)

  C<-ncol(Base.temp)

  # Create a matrix to index the alleles at each locus
  maxalleles= as.numeric(max(Allele.sum[,2]))
  Base <- matrix(data=NA,ncol=maxalleles+3, nrow=L*npops)#### rows here probably have to
extend much further!
  colnames(Base)<- c("Pop","Locus","N",rep(1:max(Allele.sum[,2])))
  Base[,1]<- rep(1:npops,rep(L,npops))
  Base[,2]<-rep(1:L,npops)

  for(k in 1:npops){
   Pop.k <- Base.temp[Base.temp$Population==k,]

   for(i in 1:L){
    allele.temp <- eval(parse(text=paste("Locus",i,sep="")))# which alleles are we dealing with
    ##add up all alleles
    a.vec<-vector(mode="numeric",length=length(allele.temp))
    for (a in 1:length(allele.temp)){
     if (i<= loci.diploid){
      a.vec[a]<-sum(Pop.k[,((2*i+1):(2*i+2))]==
eval(parse(text=paste("Locus",i,"[a]",sep=""))),na.rm=T)
     } else {
      a.vec[a]<-sum(Pop.k[,(2*i+1)]==
eval(parse(text=paste("Locus",i,"[a]",sep=""))),na.rm=T)
     }#if
     if (i<= loci.diploid){
      Base[(k-1)*L+i,4:(3+length(allele.temp))]<-a.vec # need to be clever about the 1 in the
base
      Base[(k-1)*L+i,3]<- sum(Base[(k-1)*L+i,4:(3+length(allele.temp))],na.rm=T)
     } else {
      Base[(k-1)*L+i,4:(3+length(allele.temp))]<-a.vec # need to be clever about the 1 in the
base
      Base[(k-1)*L+i,3]<- sum(Base[(k-1)*L+i,4:(3+length(allele.temp))],na.rm=T)*2
     }#if
    }#over allele
   }#over loci
```

```
   }#over pops

   #write the Base to a file
   write.table(Base,quote=FALSE,file=paste("Baseline",o,".txt",sep=""),row.names =
FALSE,col.names=FALSE, sep="\t", na="")
  }
  ##### format baseline ############################################################
  require(stringr)
  foreach (o = 1:10,.packages=c('stringr')) %dopar% {
   file.create(paste("Baseline",o,".bse",sep=""))
   line.temp<-readLines(paste("Baseline",o,".txt",sep=""))
   line.temp<-str_trim(line.temp)
   line.temp<-str_split(line.temp,"\t")
   max.ln<-npops*L
   for (l.n in 1:max.ln){
    line.temp2<-line.temp[[l.n]]
    line.temp2<-gsub(",","",paste(formatC(line.temp2, width=6,flag=" "),collapse=","))
    lapply(line.temp2, write, paste("Baseline",o,".bse",sep=""), append=TRUE)
   }}

  cat('Baseline.bse files are now saved to the working directory\n')

#################################################################################
#######
  #### Lets format all the Mixture files now
  foreach (m = 1:10,.export=Loci.names) %dopar% {
   mix.temp<- read.csv(paste("Mix",m,".csv", sep=""))

   mix <- matrix(data=NA, ncol=sum(as.numeric(Allele.sum[,2]))+(L-1),
nrow=nrow(mix.temp))

   #count alleles at each locus
   for (m.l in 1:L){
    allele.temp2 <- eval(parse(text=paste("Locus",m.l,sep="")))

    mix.a.mtrx<-matrix(data=NA,nrow=nrow(mix.temp),ncol=length(allele.temp2))

    if (m.l<= loci.diploid){ # count over locus
     for (ind.m in 1:nrow(mix.temp)){ # count for each individual in mix
      for (a.m in 1:length(allele.temp2)){ # count over all alleles at a locus
       mix.a.mtrx[ind.m,a.m]<-sum(mix.temp[ind.m,
((m.l*2+1):(m.l*2+2))]==allele.temp2[a.m],na.rm=T)
      }}
     if (m.l<2){# this section puts the results in the right place
      mix[,1:length(allele.temp2)]<- mix.a.mtrx
     } else {
```

```
        mix[,((sum(as.numeric(Allele.sum[((m.l-1):1),2]))+1)+m.l-
1):(((sum(as.numeric(Allele.sum[((m.l-1):1),2])))+(length(allele.temp2-1)))+m.l-1)]<-mix.a.mtrx
       }# if it is the first locus it needs to be placed in the right spot.
     } else { # if it is haploid
       for (ind.m in 1:nrow(mix.temp)){
         for (a.m in 1:length(allele.temp2)){
           mix.a.mtrx[ind.m,a.m]<-sum(mix.temp[ind.m, (m.l*2+1)]==allele.temp2[a.m],na.rm=T)
         }}
       mix[,((sum(as.numeric(Allele.sum[((m.l-
1):1),2])))+m.l):(((sum(as.numeric(Allele.sum[((m.l-1):1),2])))+(length(allele.temp2-1)))+m.l-
1)]<-mix.a.mtrx
       gsub("NA"," ",mix)
     }
   }
   # write the mixture to a file
   write.table(mix,quote=FALSE,na=" ",file=paste("Mixture",m,".mix", sep=""),row.names =
FALSE, sep="", col.names=FALSE)

  }
  cat ('Mixture files are now saved to working directory. Have fun running Bayes.exe')
}
##############################################################################
Ctl_file<-function(MCMC.num=1000,TI.SP=10,TI.BAF=10,TI.SAI=100,BAYES.options="T T
F T F T F",
                   Stock.ID='StockID.txt',Bayes.title='BayesTitle',Pops=74,
                   loci.haploid,loci.diploid){
# Let's make a control file
  # we have 10 mixtures and 10 baselines!

  #User input for making the control file
  #MCMC.num<-How many MCMC samples?
  #TI.SP<-thinning interval for stock proportion
  #TI.BAF<-thinning interval baseline allele freq
  #TI.SAI<-thinning interval stock assignment for individual.
  #BAYES.options<-options for output of program SEE bayes manual ex."T T F T F T F"
  #Stock.ID<- StockID block in the control file. No way to make this generic.
  #Bayes.title<- Name for the analysis
  Allele.sum<-read.csv('LociSummary.csv',header=F)
  write.table(Bayes.title,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),row.names = FALSE,
sep="", col.names=FALSE) #write name to file

write.table('Baseline1.bse',quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.
names = FALSE, sep="", col.names=FALSE)#baseline file

write.table('Mixture1.mix',quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.
names = FALSE, sep="", col.names=FALSE) #mixture file
```

```
  #output names
  Out.names<-
c('Summary.SUM','MCMC_StockProp.BOT','MCMC_AlleleFreq.FRQ','Binary_Restart.B01','St
ock_IndAssign.CLS','MCMC_SGRP.RGN')

write.table(Out.names,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.na
mes = FALSE, sep="", col.names=FALSE) #output names
  #MCMC
  write.table(format(MCMC.num,
scientific=F),quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.names =
FALSE, sep="", col.names=FALSE) #MCMC
  # number of stocks - get from input file.
  write.table(Pops,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.names =
FALSE, sep="", col.names=FALSE) #stocks
  # number of characters
  Characters<-loci.diploid+loci.haploid

write.table(Characters,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.nam
es = FALSE, sep="", col.names=FALSE) #loci
  # 3 random seeds
  Seeds<-sample(1:200000,3)
  write.table(Seeds,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.names
= FALSE, sep="", col.names=FALSE) #Seeds
  #Thinning intervals

write.table(c(TI.SP,TI.BAF,TI.SAI),quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=T
RUE,row.names = FALSE, sep="", col.names=FALSE) #Thinning intervals
  # Fortran Format for the mixture file
  Fortran.Mix<-paste('(',(paste(Allele.sum[,2],'I1,','1X',sep="",collapse=',')),')',sep="")

write.table(Fortran.Mix,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.na
mes = FALSE, sep="", col.names=FALSE) #Fortran mixture
  # Fortran Format of Baseline
  # column numbers in baseline I6)
  FtF_bse<-paste('(',(as.numeric(max(Allele.sum[,2]))+3),'I','6)',sep="")

write.table(FtF_bse,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.names
= FALSE, sep="", col.names=FALSE)

  #Options for output from BAYES

write.table(BAYES.options,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,ro
w.names = FALSE, sep="", col.names=FALSE) #Bayes Options

  # character descriptions
  Ch.Des<- matrix(data=NA,ncol=4,nrow=Characters)
```

```
  Ch.Des[,1]<-seq(from=1,to=Characters,by=1)
  Ch.Des[,2]<-Allele.sum[,2]
  Ch.Des[(1:loci.diploid),3]<-rep('T',time=loci.diploid)
  Ch.Des[((loci.diploid+1):(loci.diploid+loci.haploid)),3]<-rep('F',time=loci.haploid)
  Ch.Des[,4]<-Allele.sum[,1]
  Ch.D<-paste(formatC(Ch.Des[,1], width = 3, flag = "+"),formatC(Ch.Des[,2], width = 3, flag =
"+"),(paste(formatC(Ch.Des[,3], width = 3, flag = "+"),formatC(Ch.Des[,4], width = 3, flag = "-
"),sep=" ")),sep="")
  write.table(Ch.D,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.names =
FALSE, sep="", col.names=FALSE) #Bayes Options

  #Stock ID block
  # would this be easiest to just make the user make a stock id block?
  # read in the StockID file
  Stk.ID <- readLines(con=Stock.ID)
  #split into strings
  Stk.ID<- str_split(Stk.ID,pattern='\t')
  #what is the max length of the population name?
  C5<-max(nchar(sapply(Stk.ID, "[", 4)))
  Stk.ID.w<-paste(formatC(sapply(Stk.ID, "[", 1), width = 3, flag = "+"),formatC(sapply(Stk.ID,
"[", 2), width = 3, flag = "+"),formatC(gsub( "^0+" , "" , sapply(Stk.ID, "[", 3) ), width = 8, flag
= "+"),formatC(sapply(Stk.ID, "[", 4), width = C5+1, flag = "+"),formatC(gsub( "^0+" , "" ,
sapply(Stk.ID, "[", 5) ), width = 8, flag = "+"),sep="")

write.table(Stk.ID.w,quote=FALSE,file=paste(Bayes.title,'.ctl',sep=""),append=TRUE,row.name
s = FALSE, sep="", col.names=FALSE) #Bayes Options
  cat (paste((paste(Bayes.title,'.ctl',sep="")),'is saved in working directory',sep=" "))
}=
```

RECENT TECHNICAL MEMORANDUMS

Copies of this and other NOAA Technical Memorandums are available from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22167 (web site: *www.ntis.gov)*. Paper and electronic (.pdf) copies vary in price.

AFSC-

282 DALY, B. J., C. E. ARMISTEAD, and R. J. FOY. 2014. The 2014 eastern Bering Sea continental shelf bottom trawl survey: Results for commercial crab species, 167 p. NTIS No. PB2015-101255.

281 FAUNCE, C., J. CAHALAN, J. GASPER, T. A'MAR, S. LOWE, F. WALLACE, and R. WEBSTER. 2014. Deployment performance review of the 2013 North Pacific Groundfish and Halibut Observer Program, 74 p. NTIS No .PB2015-100579.

280 HIMES-CORNELL, A., and K. KENT. 2014. Involving fishing communities in data collection: a summary and description of the Alaska Community survey, 2010, 170 p. NTIS No PB2015-100578.

279 FISSEL, B. E. 2014. Economic indices for the North Pacific groundfish fisheries: Calculation and visualization, 47 p. NTIS No. PB2015-100577.

278 GODDARD, P., R. LAUTH, and C. ARMISTEAD. 2014. Results of the 2012 Chukchi Sea bottom trawl survey of bottomfishes, crabs, and other demersal macrofauna, 110 p. NTIS No.PB2015-100576.

277 ALLEN, B. M., and R. P. ANGLISS. Alaska marine mammal stock assessments, 2013, 294 p. NTIS No. PB2015-100575.

276 LOEFFLAD, M. R., F. R. WALLACE, J. MONDRAGON, J. WATSON, and G. A. HARRINGTON. 2014. Strategic plan for electronic monitoring and electronic reporting in the North Pacific, 52 p. NTIS No. PB2014-106286.

275 ZIMMERMANN, M., and M. M. PRESCOTT. 2014. Smooth sheet bathymetry of Cook Inlet, Alaska, 32 p. NTIS number pending.

274 ALLEN, B. M., V. T. HELKER, and L. A. JEMISON. 2014. Human-caused injury and mortality of NMFS-managed Alaska marine mammal stocks, 2007-2011, 84 p. NTIS number pending.

273 SMITH, K. R., and C. E. ARMISTEAD. 2014. Benthic invertebrates of the Eastern Bering Sea: a synopsis of the life history and ecology of the sea star *Asterias amurensis*, 60 p. NTIS number pending.

272 DE ROBERTIS, A., D. MCKELVEY, K. TAYLOR, and T. HONKALEHTO. 2014. Development of acoustic-trawl survey methods to estimate the abundance of age-0 walleye pollock in the eastern Bering Sea shelf during the Bering Arctic subarctic survey, 46 p. NTIS number pending.

271 VULSTEK, S. C., C. M. KONDZELA, C. T. MARVIN, J. WHITTLE, and J. R. GUYON. 2014. Genetic stock composition analysis of chum salmon bycatch and excluder device samples from the 2012 Bering Sea walleye pollock trawl fishery, 35 p. NTIS No. PB2014-105096.

270 GUTHRIE, C. M., III, H. T. NGUYEN, and J. R. GUYON. 2014. Genetic stock composition analysis of Chinook salmon bycatch samples from the 2012 Bering Sea and Gulf of Alaska trawl fisheries, 33 p. NTIS No. PB2014-105095.

269 MATEO, I., and D. H. HANSELMAN. 2014. A comparison of statistical methods to standardize catch-per-unit-effort of the Alaska longline sablefish, 71 p. NTIS No. PB2014-104078.

268 FOWLER, C. W., R. D. REDEKOPP, V. VISSAR, and J. OPPENHEIMER. 2014. Pattern-based control rules for fisheries management, 116 p. NTIS No. PB2014-104035.

267 FOWLER, C. W., and S. M. LUIS. 2014. We are not asking management questions, 48 p. NTIS No. PB2014-104034.