Bayesian abundance estimation from genetic mark-recapture data when not all sites are sampled: an example with the bowhead whale

Timothy R. Frasier[1], Stephen D. Petersen[2], Lianne Postma[3], Lucy Johnson[2], Mads Peter Heide-Jørgensen[4] & Steven H. Ferguson[3]

[1]Department of Biology, Saint Mary's University, 923 Robie St., Halifax, NS, B3H 3C3 Canada,
[2]Conservation and Research Department, Assiniboine Park Zoo, 2595 Roblin Blvd., Winnipeg Manitoba, R3R 0B8, Canada,
[3]Fisheries and Oceans Canada, 501 University Crescent, Winnipeg, Manitoba, R3T 2N6, Canada,
[4]Greenland Institute of Natural Resources, Box 570, 3900 Nuuk, Greenland.

Corresponding Author:
Timothy R. Frasier
Saint Mary's University
Department of Biology
923 Robie Street
Halifax, NS B3H 3C3
Canada
E-mail: timothy.frasier@smu.ca
Tel: (902) 491-6382
Fax: (902) 420-5046

Abstract

Estimating abundance is one of the most fundamental and important aspects of population biology, with major implications on how the status of a population is perceived and thus on conservation and management efforts. Although typically based on one of two methods (distance sampling or mark-recapture), there are many individual identification methods that can be used for mark-recapture purposes. In recent years, the use of genetic data for individual identification and abundance estimation through mark-recapture analyses have increased, and in some situations such genetic identifications are more efficient than their field-based counterparts for population monitoring. One issue with mark-recapture analyses, regardless of which method of individual identification is used, is that the study area must provide adequate opportunities for "capturing" all individuals within a population. However, many populations are unevenly and widely distributed, making it unfeasible to adequately sample all necessary areas. Here we develop an analytical technique that accounts for unsampled locations, and provides a means to infer "missing" individuals from unsampled locations, and therefore obtain more accurate abundance estimates when it is not possible to sample all sites. This method is validated using simulations, and is used to estimate abundance of the Eastern Canada-West Greenland (EC-WG) bowhead whale population. Based on these analyses, the estimated size of this population is 9,089 individuals, with a 95% highest density interval of 5,107–17,079.

Keywords: abundance estimation, mark-recapture, genetic identification, whale

46

Introduction

Mark-recapture analyses have become an invaluable tool not only as a means for abundance estimation, but also for obtaining a wealth of other information on population dynamics, such as population growth rates, life history, and survival probabilities [1-3]. Traditionally, mark-recapture methods involved the actual capture, marking, and recapture of individuals [4-6], but their use quickly expanded to species where individuals could be identified based on natural markings, and thus "capturing" in these cases meant taking a picture that could be used for identification [7-10]. More recently, the increasing ease with which individual-specific genetic profiles can be obtained from small sources of DNA has made it possible (and sometimes more efficient) to use genetic data as the primary means of individual identification, and thus the basis for mark-recapture analyses [11-13].

One important requirement in mark-recapture studies, regardless of which method of identification is used, is ensuring that the study area provides adequate opportunities for "capturing" all individuals within the study population. This requirement may be easily addressed for populations with small ranges that can be surveyed in their entirety, or where all individuals utilize—or pass through—a single area. However, many populations are not so amenable, with individuals exhibiting substantial variation in site fidelity and habitat use patterns. Failure to account for this variation can bias abundance estimates either upwards or downwards. The former can occur if marked individuals move out of the study area and thus become unavailable for recapture. The latter can occur if the study area is only used by a portion of the population, and thus estimates are only representative of some individuals rather than the entire population.

For species with wide distributions and substantial heterogeneity in habitat use patterns it may not be possible to obtain adequate coverage of all sites. In such cases abundance estimation requires appropriate corrections during the analytical stages of the study. In 2005, Durban et al. made substantial progress in this area by recognizing that analytically accounting for missed sites is essentially a regression problem, where data from sampled areas can be used to make predictions on the number of individuals not seen in any surveyed area [14]. As an example, suppose a study population of organisms where some, but not all, individuals can be identified based on natural markings. Suppose the number of identifiable individuals has been counted at each of three sites, but other unstudied sites exist. To estimate the total size of the identifiable (marked) population, Durban et al. first organized the sighting histories of identifiable individuals based on whether or not they were observed in each of the three surveyed areas (see **Table 1**). The last row of such a table represents the identifiable individuals not seen in *any* location, and can be inferred from the available data. In this case a regression model is fit with the first seven rows of data, and the coefficient estimates from this process are then used to predict the value for the last row. In their application of this approach to a study of bottlenose dolphins (*Tursiops truncatus*) around Scotland, Durban et al. found that their approach performed quite well and resulted in a similar estimate to that obtained through another study based on a more rigorous and standardized survey design.

The eastern Canada-western Greenland (EC-WG) bowhead whale (*Balaena mysticetus*) (**Figure 1**) population epitomizes the confounding factors associated with abundance estimation, using either distance sampling or mark-recapture methods. First, it is an arctic-adapted population with a wide distribution often associated with the ice edge

93 (**Figure 2**) [15-18]. The necessary logistics and monetary requirements make extensive
94 surveys in the high arctic challenging. Second, the population is highly segregated based
95 on sex, age class, and perhaps other factors [15,16,19,20]. Indeed, the segregation is so
96 great that, until recently, multiple populations (and subsequent management stocks) were
97 thought to exist [21]. It was only when satellite tagging data recently showed movement
98 of whales across previously assumed population boundaries, and such movements were
99 also identified based on genetic mark-recapture analyses, that the presence of only one
100 population was recognized [22-24]. These confounding factors have led to very imprecise
101 abundance estimates, as well as large differences in estimates based on data from
102 different locations and/or different analytical approaches, as summarized in Koski et al.
103 [25].

104 Here, we outline and utilize two new approaches that may improve abundance
105 estimates for the eastern Arctic-western Greenland bowhead whale population. First, our
106 analyses are based on genetic identifications. This approach may prove more efficient
107 than photograph-based identifications because samples can be collected in collaboration
108 with Arctic communities, eliminating the need for extensive and expensive aerial surveys
109 while also fostering closer collaborations and relationships with Arctic communities.
110 Second, we extend the approach of Durban et al. [14] in an important way. The model of
111 Durban et al. was designed to estimate the number of "missing" *marked* individuals in a
112 population. However, our goal was to estimate the *total number* of missing individuals
113 (marked and unmarked), and combine this information with data from sampled sites to
114 result in one estimate of total abundance for the population of interest. If successful, this
115 latter approach could prove useful to a broad range of population studies.

Materials and Methods

Sampling

Intensive sampling effort has taken place out of the communities of Igloolik and Pangnirtung in Canada, and in Disko Bay, Greenland (hereafter "Greenland") (**Figure 2**, **Table S1**). Additionally, more opportunistic sample collection has occurred in Repulse Bay and Arctic Bay, as well as from 11 subsistence hunts by Canadian Inuit (Taloyoak, Kugaaruk, Cape Dorset, Pangnirtung, Repulse Bay, Igloolik) (**Figure 2**, **Table S1**). In total, 1,177 samples have been collected across 19 years.

In each location, small tissue (skin and blubber) samples were obtained from free-swimming whales using a 150 lb draw-weight crossbow (Excalibur Vixen) with bolts and tips modified specifically for this purpose (from Ceta-DART, Denmark). This is the most common method for collecting small skin samples from free-swimming whales, and has no negative short or long-term effects on individuals other than an initial "startle" response in some cases [26-28]. Biopsy samples were collected in association with photo-identification data, as part of ongoing photo-identification studies in all areas. Samples were preserved in either a salt-saturated 20% DMSO solution [29], flash frozen in liquid nitrogen, in RNAlater (Qiagen), or in Allprotect (Qiagen) and frozen upon arrival in the laboratory. All samples were archived in -80°C freezers at the Freshwater Institute (DFO, Canada, Winnipeg) laboratories until further analysis. Bowhead whale biopsy samples were collected under permit Department of Fisheries and Oceans License to Fish for Scientific Purposes S-16/17 1005-NU and Animal Use Protocol FWI-ACC-2016-09.

138

139 ## Genetic Analyses

140 DNA was extracted using a modified version of the standard Qiagen extraction protocol
141 [30]. DNA was quantified using a µQuant spectrophotometer (Bio-Tek) or a NanoDrop
142 spectrophotemeter (ThermoFisher Scientific) and normalized to 10–100 ng/µl. Sex was
143 determined through the amplification of a zinc finger gene intron using primers LGL331
144 and LGL335 [31]. PCR fragments were size-separated in 1.5% agarose gels and
145 visualized with GelRed (Biotium). The resulting banding pattern of X (~ 975 bp) and Y
146 (~ 1040 bp) fragments were used to infer sex.

147 Multiplex reactions were developed and used to amplify 23 microsatellite loci
148 (**Table S2**). The reaction mixtures contained 1.0 µl of template DNA, 1X AmpliTaq Gold
149 Buffer, 2.0 mM $MgCl_2$, 0.25 mM each dNTP, 0.5 U AmpliTaq Gold DNA polymerase
150 (Life Technologies), and the primer concentrations indicated in **Table S2**. Thermal cycler
151 profiles included a five minute initial denaturing step at 95°C followed by 35 cycles of:
152 95°C for 30 seconds, annealing temperature (**Table S2**) for 30 seconds, and 72°C for 30
153 seconds; followed by a final extension step of 72°C for 30 minutes. PCR products were
154 size-separated and visualized on an Applied Biosystems 3130xl Genetic Analyzer (Life
155 Technologies), with GeneScan 600 LIZ (Life Technologies) as a size-standard. Size-
156 standards were checked, and allele calls made using the program GeneMarker
157 (SoftGenetics).

158

159 ## Identifying Recaptures

160 The crux in identifying genetic recaptures is developing appropriate criteria for defining
161 when two genotypes can be considered "the same", and thus representing the same
162 individual. This problem has a long history, and has been most thoroughly dealt with in
163 human forensic DNA typing [32]. The issues are two-fold. First, two genotypes may be
164 the same at the typed loci, but represent different individuals if not enough resolution is
165 available with the chosen markers, if individuals are related, or just by chance (a false
166 inclusion). Second, two genotypes may be from the same individual but be scored
167 differently due to genotyping errors or null alleles (a false exclusion). Thus, the key is
168 identifying what criteria minimize the chances of each of these errors.

169 To identify recaptures, or samples likely originating from the same individual, we
170 used the R package ALLELEMATCH [33]. This package deals with these issues by
171 calculating the genetic similarity of all genotypes, and then creating clusters based on
172 genetic similarity. It then re-creates these clusters by sequentially allowing an
173 incremental number of mismatches, while keeping track of how many "unique"
174 genotypes are present under each number of allowed mismatches. The appropriate
175 number of allowed mismatches is then identified as the point when each genotype
176 associates with only one cluster.

177

178 **Estimating Abundance**

179 We used Bayesian methods to estimate abundance from the genetic mark-recapture data,
180 based on the general approaches described in Kéry & Schaub [34] and Kruschke [35].
181 We tried two general strategies, each with their own strengths and weaknesses. The first

182 and most simple approach involved ignoring the location information and treating the
183 data as one large capture-mark-recapture study (hereafter called the "location-
184 independent" approach). The main benefit of this approach is that it allows all samples to
185 be used, regardless of whether recaptures were found within each location. The main
186 weakness of this approach is that it does not explicitly account for the fact that not all
187 locations have been sampled, and therefore makes the assumption that the sampled
188 locations represent adequate sampling opportunities for all individuals. Our second
189 strategy was modelled after that of Durban et al. [14]. Here, location-specific estimates of
190 abundance and movement rates were obtained and used to explicitly infer the number of
191 individuals in unsampled locations (hereafter called the "location-specific" approach).
192 The strength of this approach is that it provides an explicit and quantitative means to infer
193 the number of individuals using unsampled locations. The main weakness is that it
194 requires dividing the data by locations, with the consequence that many locations may not
195 have enough samples to be included in the analyses, and thus some information may be
196 lost. Both of these methods are described in more detail below.

197 The number of recaptures for this data set was small, and no individual was
198 sampled more than twice (see Results). We therefore estimated abundance using closed
199 population models because we did not feel that enough information was available to
200 obtain useful estimates of all the parameters included in an open population model.
201 Additionally, the long life-span and generally slow life history of bowhead whales mean
202 that it is unlikely that there were dramatic changes to population size throughout the
203 study period [36-39]. However, to try to account for any potential changes over time, for
204 each strategy (location-independent and location-specific) we estimated population sizes

205 based on two data sets: the full data set, and a subset containing only five years of data.
206 The rationale is that the population may have changed throughout the 19 years through
207 which the samples were collected. Therefore, closed-population abundance estimates
208 based on the entire study period may be biased. The 5-year data set contains fewer data,
209 but is also expected to be less biased by changes in population size within the sampling
210 period. Combined, this resulted in four estimates of abundance based on: (1) location-
211 independent full data set (LI-FD); (2) location-independent 5-year data set (LI-5Y); (3)
212 location-specific full data set (LS-FD); and (4) location-specific 5-year data set (LS-5Y).

213 For all analyses, the recaptures identified by ALLELEMATCH were filtered to
214 exclude recaptures from the same location in the same year. Bowhead whales are known
215 to vary their migration and habitat use patterns depending on age, sex, and life-history
216 stage [16,20]. Additionally, once in an area, whales are likely to remain there for some
217 period of residency [19]. Thus, we thought that within-year recaptures within locations
218 would be artificially high and would not be representative of general capture probabilities
219 for the entire population throughout the study period.

220

221 **Location-Independent Full Data Set (LI-FD)**

222 To estimate abundance for the entire data set, the recapture data from ALLELEMATCH
223 were first converted to typical sighting histories consisting of ones and zeros for each
224 sampling period (year) where individuals were, or were not, sampled, respectively. These
225 sighting histories were augmented by 30,000 to ensure that considered probabilities
226 covered a wide enough range to ensure adequate sampling of the posterior distribution

227 [40, 41]. The parameters of interest are the sighting probability (p), and the inclusion
228 probability (Ω), which is the probability that a member of the augmented data set is part
229 of the "true" population size (N). Abundance estimates were then obtained for a closed
230 population model (model M_0 from Otis et al. [42]) using the approach described in Royle
231 et al. [41] and Kéry & Schaub [34]. The priors for both the sighting probability (p) and
232 the inclusion probability (Ω) were uniform distributions ranging from 0 to 1. All
233 Bayesian analyses were conducted using a combination of R [43], runjags [44], and
234 JAGS [45]. The performance of models for all analyses were assessed based on
235 examination of trace plots, autocorrelation plots, the Gelman and Rubin diagnostic [46],
236 and the effective sample size, using the coda package [47].

237 **Location-Independent 5-Year Data Set (LI-5Y)**

238 Although it would make intuitive sense to select the most recent 5-year period for these
239 analyses, we did not do this because most locations did not have data for the most recent
240 year (2013). Instead, we queried the data to identify which 5-year subset would contain
241 the most sampling events. This turned out to be years 2008-2012, and therefore
242 represents a recent sampling period. Once this subset of data had been identified, an
243 estimate of abundance was obtained using the same approach as described for the full
244 data set.

245

246 **Location-Specific Full Data Set (LS-FD)**

247 Our location-specific approach for estimating abundance was modeled after that used by
248 Durban et al. [14]. The general idea is that location-specific abundance estimates can be

used in association with estimated movement rates between locations, to identify (or infer) the number of individuals with different sighting histories such as those shown in **Table 1**. When some areas are unsampled, the goal is to estimate the number of individuals in the last row (who were not captured in any area). This abundance estimate can then be used in association with estimates of movement rates to obtain an estimate of total population size even when some areas remain unsampled. The entire process is summarized in **Figure 3**, with details described below.

The first step in this process was to estimate the movement rates between all pairs of sampled locations. This is based on the number of recaptures within each considered location, and the number of those recaptures that represent individuals originally marked in the alternative location. We made the assumption that pairwise movement rates were symmetrical, largely because the number of cross-location recaptures was very low (see Results), and therefore there were few data available from which to estimate movement rates. However, it is possible to relax this assumption in the future. The number of recaptures within and between each pair of locations was converted into a string of zeros and ones, where ones represented recaptures across locations. The mode of the movement rates, and associated 95% highest density intervals (HDIs), were then estimated using a Bernoulli likelihood function and a beta prior with a uniform (flat) distribution. The result is a complete posterior probability distribution for movement rates between all pairs of sampled locations.

For the second step of the process, the probability distribution for the symmetrical movement rate to and from the unsampled location(s) is then estimated by averaging the posterior probabilities of all sampled locations. This approach makes the explicit

assumption that the rate at which individuals move between sampled and unsampled locations is within the same range as the movement rates among sampled locations. We feel that this assumption is valid in most cases, but even if not, the model still performs well overall due to a balancing effect of how the model works. Briefly, estimating the movement rates themselves are not the goal of the model, but rather they are used to estimate the abundance of individuals using different locations. Therefore, if the estimated movement rate to and from the unsampled location(s) is *lower* than the true value, then this will result in the abundance estimates for the sampled locations being *lower* than their true values (because it will be assumed that fewer individuals are missed by moving to unsampled locations), whereas the abundance estimate for the unsampled location(s) will be *higher* than its true value (because it will be assumed that fewer individuals are moving into and out of that area, and thus also captured in other areas). However, because the estimated movement rate is an average, these over- and under-estimates will balance each other out, and the overall estimate for the total population size should still be accurate. The same will be true for the opposite scenario where the estimated movement rate to and from the unsampled location(s) is *higher* than the true value.

For the third step, abundance is estimated for each sampled location. The genetic recapture data were first converted to typical sighting histories consisting of ones and zeros for each sampling period where individuals were, or were not, captured, respectively. However, these sighting histories are biased due to individuals moving to and from the unsampled location(s). Individuals marked in the sampled location could move to an unsampled location, and therefore not be available for recapture, artificially

295 reducing the number of recaptures. In the opposite direction, unmarked individuals from
296 the unsampled location(s) could move into the sampled location, which would also
297 artificially reduce the proportion of recaptures. Both of these processes reduce the
298 proportion of individuals available for recapture, and would therefore bias population size
299 estimates upwards. To correct for this, for each sampled location we estimate the number
300 of individuals moving to and from the unsampled location(s) and randomly add these to
301 the sighting history data. For example, suppose at one location there were 300 marked
302 individuals and 68 recaptures. Suppose that the estimated movement rate to and from the
303 unsampled location(s) is 0.12. This means that we have "lost" 12% of marked individuals
304 to unsampled locations, and our number of unmarked individuals is also artificially
305 inflated by 12% due to movement of unmarked individuals from unsampled areas into
306 this area. This means that our proportion of recaptures is underestimated by two times the
307 movement rate (24%). Thus, our original 68 recaptures only represent (1 - 0.24 = 0.76)
308 76% of the recaptures we "should" have had, if individuals did not move to and from
309 unsampled locations (i.e., we should have had 68 / 0.76 = 89 recaptures). To account for
310 this, these (89 - 68 = 21) additional recaptures were randomly added to the sighting
311 histories for this location. Once the sighting histories are corrected, abundance was
312 estimated for each sampled location using the same closed population approach described
313 for the location-independent data sets. Note that although the model assumed a closed
314 population, movement to and from other locations has already been taken into account,
315 and therefore "closed" in this sense just refers to births and deaths.

316 For step four, the complete posterior probabilities for abundance for each sampled
317 location, and for each pairwise movement rate, are used to estimate the number of

individuals with each possible habitat use pattern across the sampled locations (there are 3 such possible patterns for 2 locations - **Figure 3**, 7 possible patterns for 3 locations, and so on). These data are then used in a Bayesian regression analysis to estimate the relevant regression coefficients. These coefficients are then used to predict the number of individuals that are not seen in any of the sampled locations (i.e., the last row of the table shown in **Figure 3**, step 4). The resulting posterior probability distribution represents the estimate of the total number of individuals not seen in any sampled location.

For the fifth and final step, total abundance is estimated in a similar manner as that for the "missing" individuals. Specifically, the complete posterior probabilities for abundance for each location, and for each pairwise movement rate, are used to estimate the number of individuals with each possible habitat use pattern across locations. The difference is that now this table includes data for the unsampled location(s): effectively adding another location to the table and associated movement rates (see **Figure 3**). The sum of these values then represents the abundance estimate for the total population. Note that one benefit of this approach is that the uncertainty in each estimate is carried over into the uncertainty of the total abundance estimate. Put another way, rather than estimating the total abundance based on the mean, mode, or some other summary metric of the other parameters, this estimate is based on the entire probability distribution for each parameter. In this way the uncertainty in each estimate of abundance and movement are incorporated into the posterior probability distribution for the total abundance estimate. This is one of the attractive properties of setting up the analyses in this way.

Location-Specific 5-Year Data Set (LS-5Y)

As with the location-independent 5-year data set, we conducted the same analyses as above (for the location-specific full data set, LS-FD), except this time with a data set containing only those sampling events in the 5-year period from 2008-2012. Again, the idea for this process was to reduce the potential biases in the longer-term data set due to births and deaths occurring within the longer time frame, by analyzing data from a much shorter time period.

Model Validation

A number of simulations were conducted to assess model performance, and to provide information that would be useful for guiding sampling strategies in the field. All simulations were based on the location-specific model, which, instead of estimating abundance for a single population/location, is designed to estimate abundance for each location independently (including unsampled locations), and then come up with one total abundance estimate for all sampled and unsampled locations. A simulation function was written that generates data for four sampled locations and one unsampled location, where the abundance and movement rates between locations/populations can be set by the user.

The first set of simulations focused on assessing the effects of varying sampling effort. For these simulations all population sizes (including that of the unsampled location) were set to 1,000 individuals (resulting in a total population size of 5,000 individuals). All movement rates between locations were also set to 10%. Abundance was then estimated based on sampling 5%, 10%, 15%, 20%, and 25% of individuals from

362 each sampled location, to assess how changing sampling effort influences the precision
363 and accuracy of abundance estimates.

364 A second set of 5 simulations were also run and analyzed, but this time the
365 abundance at each location (including the unsampled one) was randomly drawn from a
366 list of 20 values ranging from 500 to 10,000, in increments of 500. The movement rates
367 were all set to 10%, and 250 individuals were sampled at each location. This means that
368 the proportion of individuals sampled could range from 2.5% (for a population size of
369 10,000 individuals) up to 50% (for a population size of 500 individuals). These
370 simulations were conducted to ensure that the model performs well even when there are
371 large differences in population sizes and sampling effort between locations.

372

373 # Results

374 ## Samples, Genotypes, and Recaptures

375 A total of 1,177 samples were genotyped. **Table 2** shows how these were divided among
376 locations. The final microsatellite data set contained 21 loci since two loci were dropped
377 due to a significant portion of the samples missing data that were profiled before 2008.
378 Furthermore, a sample was only included in the analysis if it had at least 10 of the 21 loci
379 scored. This provided an overall probability of identity of 1.16×10^{-9}, which corresponds
380 to a 1 in 8.62×10^{8} chance that the match identified was due to two random animals
381 matching at all loci. **Supplementary File 1** shows how these samples were divided by
382 year, location, and sex.

383 The ALLELEMATCH analysis suggested that the best fit for the data was to allow 3
384 mismatches. Based on this criterion, 992 unique genotypes were identified, along with
385 185 recaptures. However, 136 of these recaptures represented whales recaptured within
386 the same area and year in which they were marked, and were therefore removed from the
387 analyses. This left 49 recaptures for subsequent analyses. **Table 3** summarizes the
388 recaptures by location. Based on the number of recaptures, only Greenland, Igloolik,
389 Pangnirtung, and Repulse Bay were used for estimating abundance based on the full data
390 set.

391

392 ## Model Validation

393 The results for the simulations testing the effects of sampling effort are shown in **Figure**
394 **4**. The precision of the estimates (or the range between to upper and lower 95% HDI
395 values) declines rapidly until sampling reaches ~15%. Above this sampling effort,
396 precision increases at a much slower rate. Thus, ~15% seems to be a turning point above
397 which diminishing returns are achieved as sampling effort increases. The details of all of
398 the estimates from these simulations are shown in **Figure 5**. These plots show that the
399 model performs quite well in all aspects (abundance estimation of sampled locations,
400 abundance estimation of unsampled location(s), and total abundance).

401 The results from the simulations testing the performance when population size
402 and sampling effort varies across locations are shown in **Figure 6**. The results show that
403 the model performs quite well under a range of variation in population sizes and sampling
404 efforts. This is particularly true for the estimates of "total" abundance, which tend to be

more accurate than the estimate for each specific location. This is due to differences between the true and estimated movement rates as described in the Methods section.

It should be noted that all of the simulation results are based on only one iteration of the simulation under each condition. Ideally, simulations would be run 10, 100, or 1000 times under each set of conditions to obtain information about the variance around expectations under each conditions. However, given the way the model works, it takes approximately one week to analyze the data from one simulation. Therefore, conducting and analyzing a more desirable number simulations would take a prohibitive amount of time. As a result, these results should be taken as a representative, but not a comprehensive, evaluation of model performance under the simulated conditions.

Bowhead Whale Abundance Estimates

The location-specific abundance estimates for each location, inferred for the unsampled location(s), and for the total population, are shown in **Table 4** (rows 1–6). The mode was used as the point estimate instead of the mean in all cases because the posterior probability distributions for all abundance estimates were skewed, with heavy right-hand tails, and therefore the mode represented the estimate with a higher probability than the mean. Note that the estimates for each location are inclusive, meaning that they represent *all* whales that use each area (and may also be counted for other areas), rather than representing whales that *only* use each area. The location–independent abundance estimates based are also shown in **Table 4** (row 7).

Discussion

Analyses of simulated data sets suggest that our new approach performs quite well, and could be a useful tool for estimating abundance of other populations where only a subset of possible locations have been sampled. Simulations testing analysis performance with respect to sampling effort show that sampling roughly 15% of the individuals at each location is desirable, whereas sampling more intensively will result in diminishing returns. The approach still performs well when fewer than 15% of individuals are sampled at each location, but there is decreasing precision around the estimates as sample size decreases (**Figure 5**).

The approach also performs well even when there is wide variation in abundance between sampling locations. Although the simulations used in our tests were not exhaustive, they do show that accurate abundance estimates can be obtained for the total population when only a subset of locations are sampled, and when there is wide variation in abundance across locations (**Figure 6**). Combined, the analysis of simulated scenarios validates this new approach, and shows that it is indeed a viable option for estimating abundance from genetic mark-recapture data when not all locations are sampled.

For analyses of the bowhead whale data specifically, the number of recaptures was very low, despite the substantial effort in sample collection across a wide range of areas. This fact alone suggests that the population size is quite large, and demonstrates the inherent difficulties in estimating abundance for EC-WG bowhead whales. Given their movement dynamics and heterogeneity, it seems that any estimate of population size

is going to have its shortcomings, and have to be taken with a fairly large grain of salt. However, here we have tried to be comprehensive in our analyses to try to make the most of the data that are available.

For the location-specific estimates for sampled locations, our results are comparable to previously published estimates that are available. First, Cosens & Innes [48] report abundance estimates from aerial line-transect surveys for a larger area centred around Repulse Bay. Their estimate was 75 individuals, with a 95% confidence interval of 17–133. This agrees well with our estimate of 30 (95% HDI 19–71), particularly when considering that they covered a larger area than was available here. Second, Rekdal et al. [49] report aerial survey line-transect and genetic mark-recapture abundance estimates for West Greenland. The estimates from these methods were 744 whales (95% CI: 357–1,461) and 1,538 whales (827–2,249), respectively. However, the aerial surveys were conducted in a single year, and therefore the resulting estimates do not include whales that may use the area, but did not do so in the year of the survey, which may represent a substantial number. Our genetic mark-recapture abundance estimate for West Greenland is (1,777, 95% HDI: 855–5,536), which is very similar to that of Rekdal et al. [49], but with a heavier right-handed tail. This makes sense given that our approach attempts to infer whales that may be missing, and therefore results in higher probabilities associated with larger estimates. Given the heterogeneity in bowhead movement patterns, we think that basing estimates on a longer-term data set increases the chances of including all whales that use specific areas, and will therefore result in more accurate estimates.

For total abundance estimates, our values were fairly similar across approaches (location-independent and location-dependent), as well as across time periods used (full

472 data set versus 5 years of data) (**Table 4**). In most cases the estimates for the 5-year data
473 set were slightly lower than those for the whole data set. This could result for two
474 primary reasons. First, the higher estimates may be correct, but only obtained with the
475 larger data set due to the much larger sample size available. Second, the higher estimates
476 may be biased by population growth during the study period, which would result in
477 abundance estimates for the longer time period being biased upwards. Precision was
478 higher for the longer time period, which was expected due to the larger number of
479 recaptures available for the longer-term data set. Given these differences and potential
480 issues, we think that the location-dependent estimate based on the 5-year data set
481 currently represents the best overall estimate from these analyses (9,089 whales, 95%
482 HDI: 5,107–17,079). This estimate is higher than that based on the location-independent
483 approach, but this is expected because the location-independent approach cannot account
484 for "missing" whales that do not utilize unsampled areas. The precision of this estimate is
485 quite low, indicating that effort should be improved in the future to obtain samples from
486 these locations. Interestingly, Fisheries and Oceans Canada (DFO) recently (in 2013)
487 carried out large-scale aerial surveys of the arctic to estimate abundance of narwhal and
488 bowhead whale populations. The estimate they obtained for the EC-WG bowhead
489 population was 6,743 individuals (CV = 22%) [50]. However, this is likely an
490 underestimate given that several key areas were not covered by those surveys (Foxe
491 Basin, and northern Hudson Bay). Thus, we would expect our estimate to be larger than
492 that reported based on the aerial surveys.

493 These analyses show that, at least in the tested conditions, it is possible to use
494 regression techniques to infer "missing" individuals in mark-recapture abundance

analyses, and therefore improve abundance estimates for situations where it is not possible to sample all locations. This advancement should be useful to a wide range of studies, and although it has its limitations and issues, and is by no means a final solution, it does represent a step forward in the quest to obtain accurate estimates of abundance for populations that are difficult to study. For studies and management of the EC-WG bowhead whale population, the benefits of this approach may take many other forms as well. For example, the systematic collection and analysis of biopsy samples from a few key locations is vastly less expensive than conducting exhaustive aerial surveys of the high arctic. Thus, this may represent a more efficient means to assess and monitor this population over time than the methods that are currently employed.

Data Availability

The data, code, and instructions for combining the two to obtain the results are available at https://github.com/timothyfrasier/bowhead_abundance. Specifically, there is a pdf file containing "walk-through" instructions, explaining how to conduct the analyses and obtain the results presented in the paper. There are also two folders: one containing the required data, and another containing the required code. Users should be able to replicate our results by following the instructions.

Acknowledgements

Over the years numerous people have been involved in sample collection. This work would not have been possible without their participation. We sincerely thank the Hunters

and Trappers Organizations and the communities of Igloolik, Pangnirtung, Arctic Bay, Repulse Bay, Kuugaaruk, Taloyoak and Cape Dorset who provided logistic support, field assistance and/or participation for the biopsy sampling and collection programs. We appreciate the work done by Denise Tenkula (DFO) for laboratory analyses and scoring of alleles. We are grateful to DFO staff for their efforts in receiving, handling and curating of samples. Funding for the sample collection and genetic analyses has been obtained from DFO, Polar Continental Shelf Project, Nunavut Implementation Fund, the Nunavut Wildlife Research Trust, ArcticNet Centre of Excellence, International Polar Year (Global Warming and Arctic Marine Mammals), Assiniboine Park Zoo, the University of Manitoba and Natural Sciences and Engineering Research Council. This work was funded by Fisheries and Oceans Canada (DFO).

References

1. Kraus SD, Hamilton PK, Kenney RD, Knowlton AR, Slay CK. Reproductive parameters of the North Atlantic right whale. J. Cetacean Res. Manag. 2001; Special Issue 2: 231-236.

2. Williams BK, Nichols JD, Conroy MJ. Analysis and Management of Animal Populations. Sand Diego: Academic Press; 2002.

3. Amstrup SC, McDonald TL, Manly BFJ. Handbook of Capture-Recapture Analysis. Princeton: Princeton University Press; 2005.

4. Rice WR, Harder JD. Application of multiple aerial sampling to a mark-recapture census of white-tailed deer. J. Wildl. Manage. 1977;41: 197-206.

5. Hestbeck JB, Nichols JD, Malecki RA. Estimates of movement and site fidelity using mark-resight data of wintering Canada geese. Ecol. Evol. 1991;72: 523-533.

6. Pradel R, Johnson AR, Viallefont A, Nager RG, Cézilly F. Local recruitment in the greater flamingo: a new approach using capture-mark-recapture data. Ecol. Evol. 1997;78: 1431-1445.

7. Wilson B, Hammond PS, Thompson PM. Estimating size and assessing trends in a coastal bottlenose dolphin population. Ecol. Appl. 1999;9: 288-300.

8. Matthews JN, Steiner L, Gordon J. Mark-recapture analysis of sperm whale (*Physeter macrocephalus*) photo-id data from the Azores (1987-1995). J. Hered. 2001;3: 219-226.

9. Langtimm CA, Beck CA, Edwards HH, Fick-Child KJ, Ackerman BB, Barton SL, et al. Survival estimates for Florida manatees from the photo-identification of individuals. Mar. Mamm. Sci. 2004;20: 438-463.

10. Bradshaw CJA, Mollet HF, Meekan MG. Inferring population trends for the world's largest fish from mark-recapture estimates of survival. J. Anim. Ecol. 2007;76: 480-489.

11. Mowat G, Strobeck C. Estimating population size of grizzly bears using hair capture, DNA profiling, and mark-recapture analysis. J. Wildl. Manage. 2000;64: 183-193.

12. Miller CR, Joyce P, Waits LP. A new method for estimating the size of small populations from genetic mark-recapture data. Mol. Ecol. 2005;14: 1991-2005.

13. Petit E, Valiere N. Estimating population size with noninvasive capture-mark-recapture data. Conserv. Biol. 2006;20: 1062-1073.

14. Durban JW, Elston DA, Ellifrit DK, Dickson, E. Hammond PS, Thompson PM. Multisite mark-recapture for cetaceans: population estimates with Bayesian model averaging. Mar. Mamm. Sci. 2005;21: 80-92.

15. Reeves R, Mitchell E, Mansfield A, McLaughlin M. Distribution and migration of the bowhead whale, *Balaena mysticetus*, in the eastern North American Arctic. Arctic 1983;36: 5-64.

16. Moore SE, Reeves RR. Distribution and movement. In: Burns JJ, Montague JJ, Cowles CJ, editors. The bowhead whale. Lawrence: The Society of Marine Mammalogy; 1993. pp. 313-386.

17. Finley KJ. Natural history and conservation of the Greenland whale, or bowhead, in the northwest Atlantic. Arctic 2001;54: 55-76.

18. Ferguson SH, Dueck L, Loseto LL, Luque SP. Bowhead whale *Balaena mysticetus* seasonal selection of sea ice. Mar. Ecol. Prog. Ser. 2010;411: 285-297.

19. Finley K. Isabella Bay, Baffin Island: an important historical and present-day concentration area for the endangered bowhead whale (*Balaena mysticetus*) of the eastern Canadian Arctic. Arctic 1990;43: 137-152.

20. Heide-Jørgensen MP, Laidre KL, Wiig Ø, Postma L, Dueck L, Bachmann L. Large-scale sexual segregation of bowhead whales. Endanger. Species Res. 2010;13: 73-78.

21. Rugh D, Demaster D, Rooney A, Breiwick J, Shelden K, Moore S. A review of bowhead whale (*Balaena mysticetus*) stock identity. J. Cetacean Res. Manag. 2003;5: 267-279.

22. Heide-Jørgensen MP, Laidre KL, Wiig Ø, Jensen MV, Dueck L, Maiers LD, et al. From Greenland to Canada in ten days: tracks of bowhead whales, *Balaena mysticetus*, across Baffin Bay. Arctic 2003;56: 21-31.

586 23. Heide-Jørgensen MP, Laidre KL, Jensen MV, Dueck L, Postma LD. Dissolving
587 stock discreteness with satellite tracking: bowhead whales in Baffin Bay. Mar.
588 Mamm. Sci. 2006;22: 34-45.

589 24. Wiig Ø, Heide-Jørgensen MP, Lindqvist C, Laidre KL, Postma LD, Dueck L, et al.
590 Recaptures of genotyped bowhead whales *Balaena mysticetus* in eastern Canada
591 and West Greenland. Endanger. Species Res. 2011;14: 235-242.

592 25. Koski WR, Da-Silva CQ, Zeh J, Reeves RR. Evaluation of the potential to use
593 capture-recapture analyses of photographs to estimate the size of the eastern
594 Canada west Greenland bowhead whale (*Balaena mysticetus*) population. Can.
595 Wild. Biol. Manage. 2013;2: 23-35.

596 26. Best P, Reeb D, Rew M, Palsbøll PJ, Schaeff C, Brandão A. Biopsying southern
597 right whales: Their reactions and effects on reproduction. J. Wildl. Manage.
598 2005;69: 1171-1180.

599 27. Noren DP, Mocklin JA. Review of cetacean biopsy techniques: Factors
600 contributing to successful sample collection and physiological and behavioral
601 impacts. Mar. Mamm. Sci. 2011;28: 154-199.

602 28. Kowarski KA, Augusto JF, Frasier TR, Whitehead H. Effects of remote biposy
603 sampling on long-finned pilot whales (*Globicephala melas*) in Nova Scotia. Aquat.
604 Mamm. 2014;40: 117-125.

605 29. Seutin G, White B, Boag P. Preservation of avian blood and tissue samples for
606 DNA analyses. Can. J. Zool. 1991;69: 82-90.

607 30. Wang JY, Frasier TR, Yang SC, White BN. Detecting recent speciation events: the
608 case of the finless porpoise (genus *Neophocaena*). Heredity. 2008;101: 145-155.

609 31. Shaw C, Wilson P, White B. A reliable molecular method of gender determination
610 for mammals. J. Mammal. 2003;84: 123-128.

611 32. National Research Council. The evaluation of forensic DNA evidence. Washington,
612 D.C.: National Academy Press; 1996.

613 33. Galpern P, Manseau M, Hettinga P, Smith K, Wilson P. Allelematch: an R package
614 for identifying unique multilocus genotypes where genotyping error and missing
615 data may be present. Mol. Ecol. Resour. 2012;12: 771-778.

616 34. Kéry M, Schaub M. Bayesian Population Analysis Using WinBUGS. Waltham:
617 Academic Press; 2012.

618 35. Kruschke JK. Doing Bayesian data analysis. Burlington: Academic Press; 2011.

619 36. George JC, Bada J, Zeh J, Scott L, Brown SE, O'Hara T, et al. Age and growth
620 estimates of bowhead whales (*Balaena mysticetus*) via aspartic acid racemization.
621 Can. J. Zool. 1999;77L 571-580.

622 37. George JC, Bockstoce JR. Two historical weapon fragments as an aid to estimating
623 the longevity and movements of bowhead whales. Polar Biol. 2008;31: 751-754.

624 38. Nerini MK, Braham HW, Marquette WM, Rugh DJ. Life history of the bowhead
625 whale, *Balaena mysticetus* (Mammalia: Cetacea). J. Zool. 1984;204: 443-468.

626 39. Zeh J, Poole D, Miller G, Koski W, Baraff L, Rugh D. Survival of bowhead
627 whales, *Balaena mysticetus*, estimated from 1981-1998 photoidentification data.
628 Biometrics 2002;58: 832-840.

629 40. Tanner MA, Wong WH. The calculation of posterior distributions by data
630 augmentation. J. Am. Stat. Assoc. 1987;82: 528-540.

631 41. Royle JA, Dorazio RM, Link WA. Analysis of multinomial models with unknown
632 index using data augmentation. J. Comput. Graph. Stat. 2007;16: 67-85.

633 42. Otis DL, Burnham KP, White GC, Anderson DR. Statistical inference from capture
634 data on closed animal populations. Wildl. Monogr. 1978;62: 3-135.

635 43. R Core Team. *R:* A language and environment for statistical computing. Vienna: R
636 Foundation for Statistical Computing; 2015.

637 44. Denwood MJ. runjags: An R package providing interface utilities, model templates,
638 parallel computing methods and additional distributions for MCMC models in
639 JAGS. J. Stat. Softw. 2015;21: 80-92.

640 45. Plummer M. JAGS: A program for analysis of Bayesian graphical models using
641 Gibbs sampling. Proceedings of the 3rd International Workshop on Distributed
642 Statistical Computing (DSC 2003), Vienna, Austria; 2003.

643 46. Gelman A, Rubin DB. Inference from iterative simulation using multiple
644 sequences. Stat. Sci. 1992;7: 457-511.

645 47. Plummer M, Best N, Cowles K, Vines K. CODA: convergence diagnosis and
646 output analysis for MCMC. R News 2006;6: 7-11.

48. Cosens SE, Innes S. Distribution and numbers of bowhead whales (*Balaena mysticetus*) in northwestern Hudson Bay in August 1995. Arctic 2000;53: 36-41.

49. Rekdal SL, Hansen RG, Borchers D, Bachmann L, Laidre KL, Wiig Ø, et al. Trends in bowhead whales in West Greenland: Aerial surveys vs. genetic capture-recapture analyses. Mar. Mamm. Sci. 2014;31: 133–154.

50. Doniol-Valcroze T, Gosselin JF, Pike D, Lawson J, Asselin N, Hedges K, et al. Abundance estimate of the Eastern Canada - West Greenland bowhead whale population based on the 2013 High Arctic Cetacean Survey. DFO Can. Sci. Advis. Sec. Res. Doc. 2014/WP 2b, Ottawa, ON.

Table 1. A table showing the data layout as needed to infer the number of marked individuals in unsampled sites, as shown in Durban et al. (2005). Each row represents individuals with a particular capture history. For example, row one represents individuals seen in all three areas. Row two represents individuals seen in areas #1 and #2, but not area #3. Row three represents individuals seen in area #1 and area #3, but not area #2, and so on. The last row represents identifiable individuals not seen in any of the surveyed areas, and the number of individuals with this sighting history is to be inferred.

Marked Individuals	**Area 1**	**Area 2**	**Area 3**
1	1	1	1
16	1	1	0
3	1	0	1
35	1	0	0
2	0	1	1
2	0	1	0
16	0	0	1
??	0	0	0

666 **Table 2**. Number of genotyped samples by location.

667

Location	Count
Arctic Bay	9
Cape Dorset	1
Greenland	295
Igloolik	546
Iqaluit	1
Kugaaruk	8
Pangnirtung	293
Repulse Bay	23
Taloyoak	1
Total	**1,177**

668

669

Table 3. Identified recaptures by location. Only locations with recaptures are included.

	Greenland	Igloolik	Kugaaruk	Pangnirtung	Repulse Bay
Greenland	5	0	0	2	0
Igloolik	-	23	1	6	0
Kugaaruk	-	-	0	0	0
Pangnirtung	-	-	-	3	3
Repulse Bay	-	-	-	-	1

673 **Table 4**. Abundance estimates for each location, as well as for the full population.
674 Included are these estimates and 95% highest density intervals (HDI). Total-LS refers to
675 the abundance estimate for the total population based on inference from the location-
676 specific analyses and Total-LI refers to estimates based on the location-independent
677 analyses.

		Full Data Set		**5 Years of Data**	
Row #	**Location**	**Estimate**	**95% HDI**	**Estimate**	**95% HDI**
1	Greenland	2,254	1,250–5,097	1,777	855–5,536
2	Igloolik	3,629	2,638–5,098	2,055	1,424–3,013
3	Pangnirtung	2,179	1,257–3,659	2,510	1,150–6,102
4	Repulse Bay	30	19–71	NA	NA
5	“Missing”	1,694	0–3,940	2,122	0–10,095
6	Total-LS	10,331	7,639–14,558	9,089	5,107–17,079
7	Total-LI	12,220	8,680–16,200	7,660	4,500–11,100

678

679

680

681

682 **Figure 1.** Bowhead whale (*Balaena mysticetus*). Photograph by Jeff Higdon.

683

684 **Figure 2**. Distribution and sampling locations of the eastern Arctic bowhead population.
685 Sampling locations are numbered as follows: 1 = Arctic Bay, 2 = Cape Dorset, 3 =
686 Igloolik, 4 = Kugaaruk, 5 = Pangnirtung, 6 = Repulse Bay, 7 = Taloyoak, and 8 = Disko
687 Bay.

688

689 **Figure 3**. Illustration and brief description of the main steps of the model.

690

691 **Figure 4**. Range of abundance estimates (interval between the upper and lower 95% HDI
692 values) obtained for total abundance when 5, 10, 15, 20, and 25% of the individuals from
693 each location are sampled.

694

695 **Figure 5**. Abundance estimates and associated 95% HDIs for each simulated scenario
696 testing the effects of sampling effort. Actual values are shown in black circles. Here
697 "Missing" refers to the number of individuals in the unsampled location(s).

698

699 **Figure 6**. Abundance estimates and associated 95% HDIs for each simulated scenario
700 testing the effects of varying population sizes and sampling efforts across locations.
701 Actual values are shown in black circles. Here "Missing" refers to the number of
702 individuals in the unsampled location(s). Note that the "total" abundance estimate for all
703 5 tested scenarios are located in the bottom-right panel. These were separated from the

other results for each simulated condition because the scale is so different for the total estimates versus those for each location independently.

Figure 1

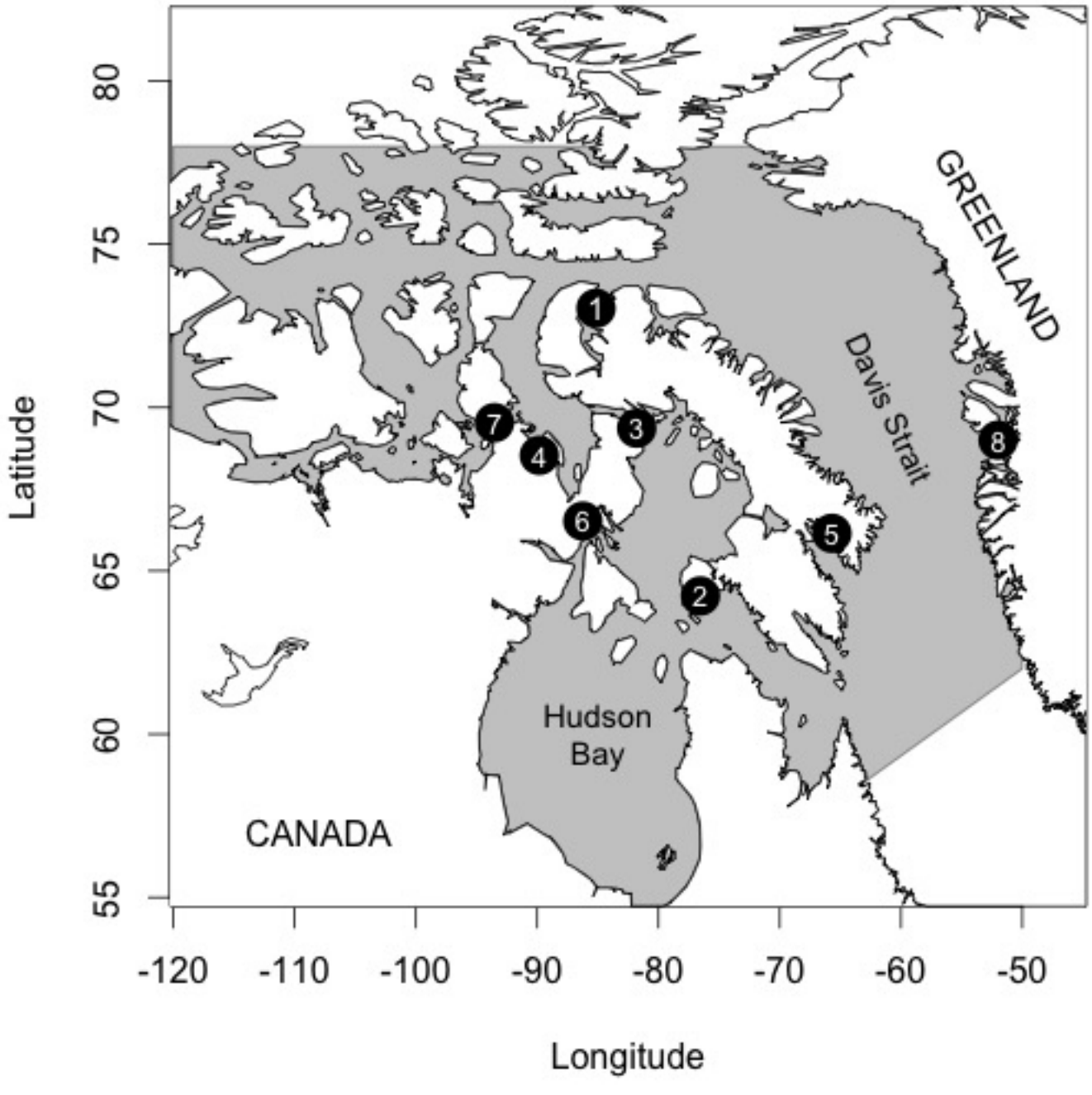
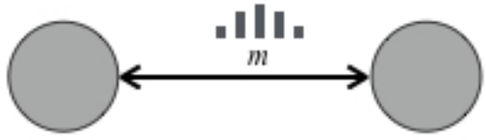

GREENLAND
Davis Strait
Hudson
Bay
CANADA
1
2
3
4
5
6
7
8
Latitude
80
75
70
65
60
55
Longitude
-120
-110
-100
-90
-80
-70
-60
-50

Figure 2

Step 1:



Estimate movement rates between all pairwise combinations of sampled locations. Assume movements are symmetrical. Bayesian estimation provides complete probability distributions for possible values.

Step 2:

Estimate movement rate to unsampled location(s) as the average across all pairwise combinations. Again, this results in a complete probability distribution for this estimated movement rate.

Step 3:

Estimate abundance for each sampled location, correcting for the movement of individuals to and from unsampled location(s). Bayesian estimation provides complete probability distributions for possible values.

Step 4:

Use Pattern	Loc. 1	Loc. 2
1	YES	YES
2	YES	NO
3	NO	YES
4	NO	NO

Use abundance estimates for each sampled location and estimated movement rates to estimate the number of individuals with all possible habitat use patterns across sampled areas (example table with just 2 locations shown). Build regression model from data, and predict the number of individuals that don't use any sampled location (last row of table). This is the estimate of "missing" individuals.

Step 5:

Use Pattern	Loc. 1	Loc. 2	Unsampled
1	YES	YES	YES
2	YES	YES	NO
3	YES	NO	YES
4	NO	YES	YES
5	YES	NO	NO
6	NO	YES	NO
7	NO	NO	YES

Use abundance estimates for all locations (including unsampled ones) and estimated movement rates to estimate the number of individuals with all possible habitat use patterns. The sum of these is the abundance estimate for the total population.

Figure 3

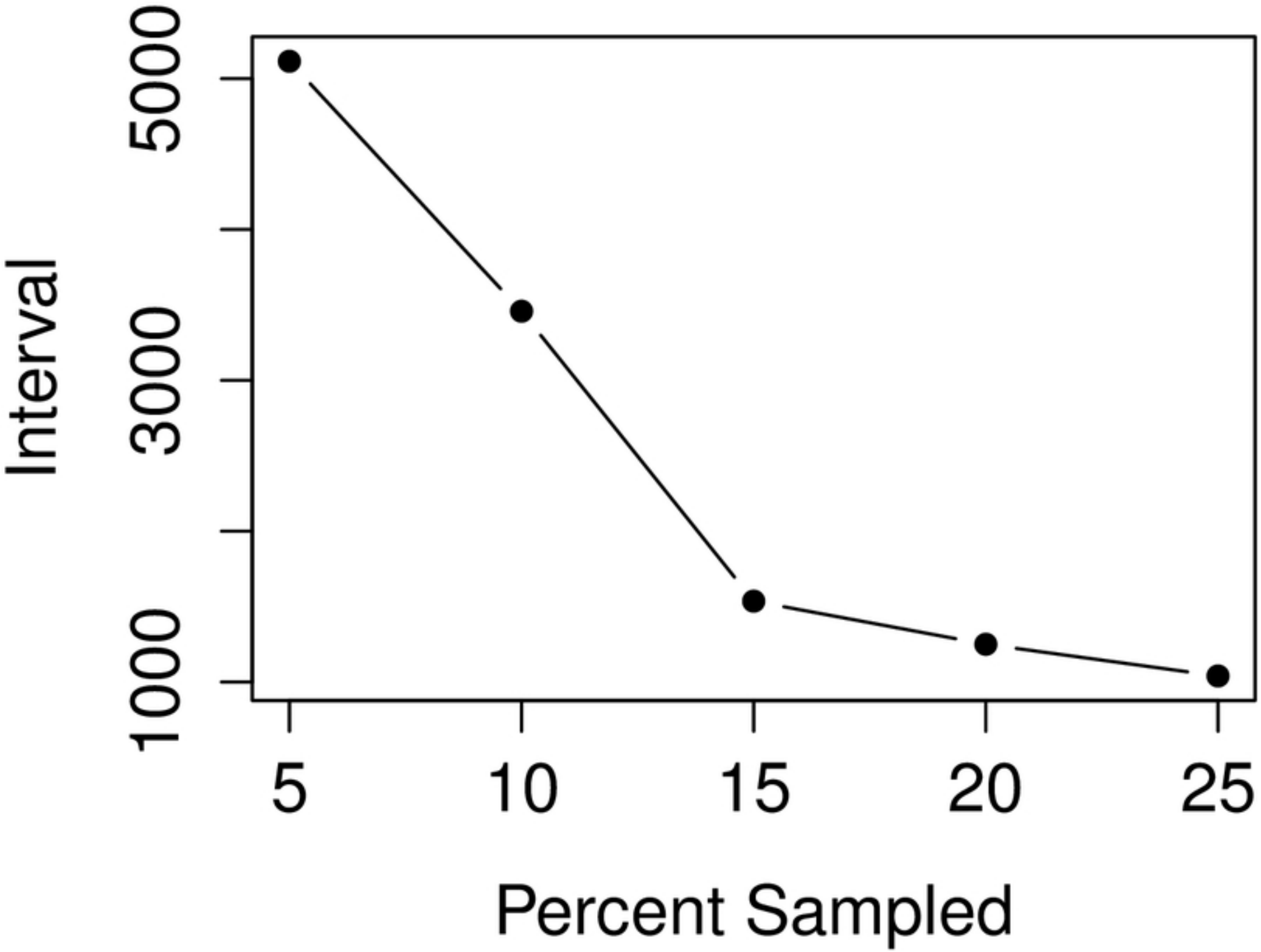
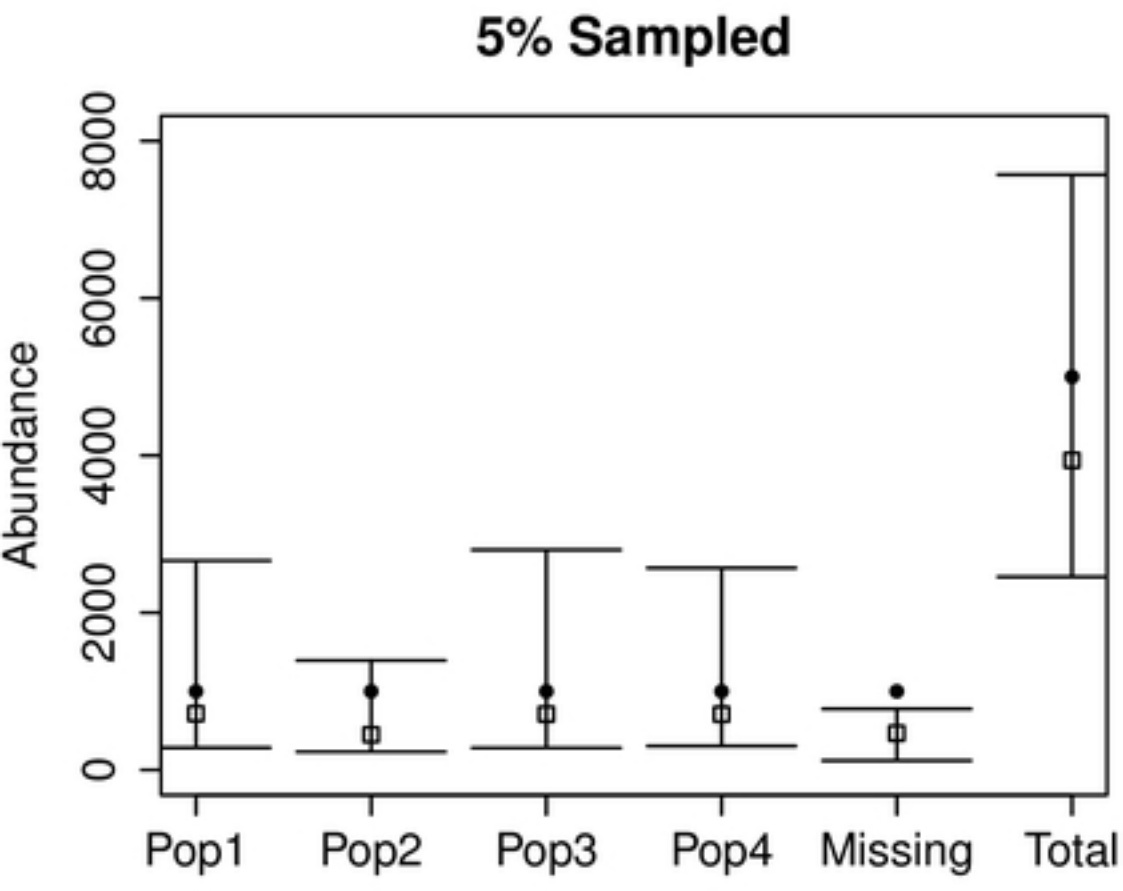
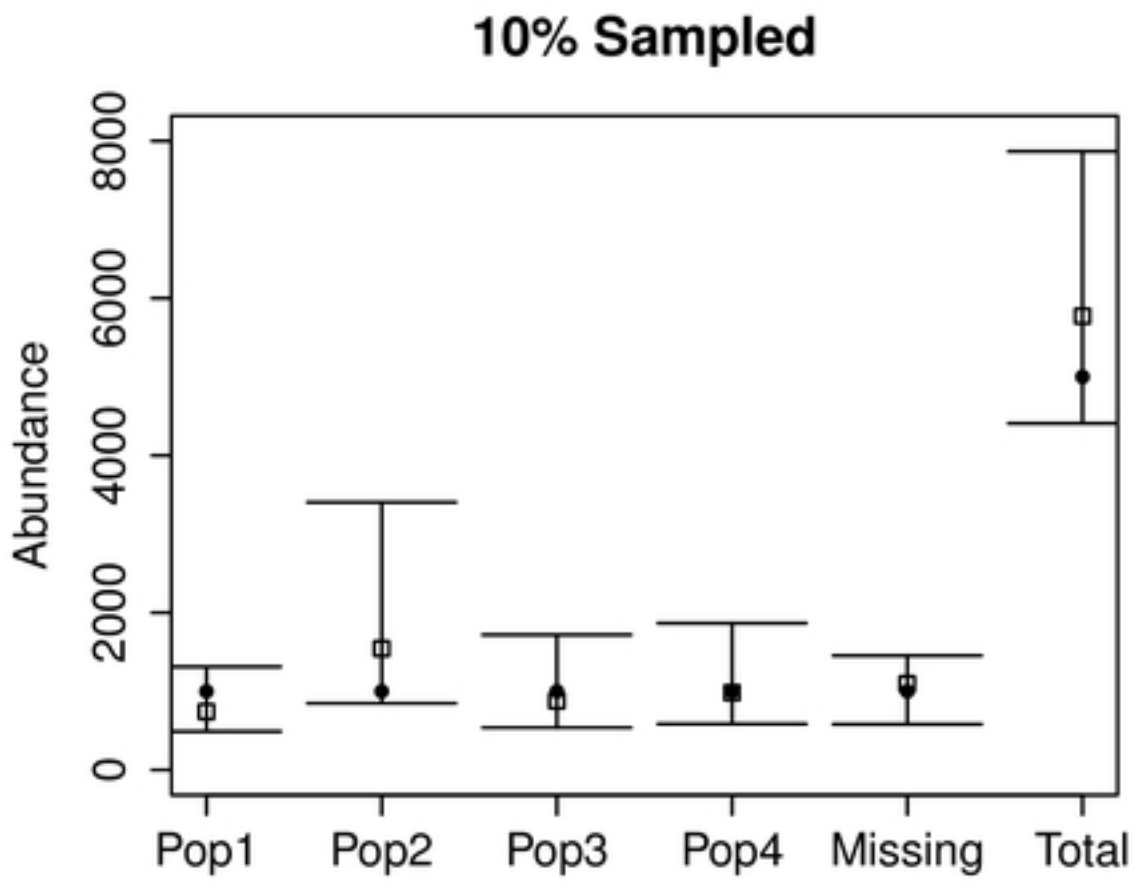
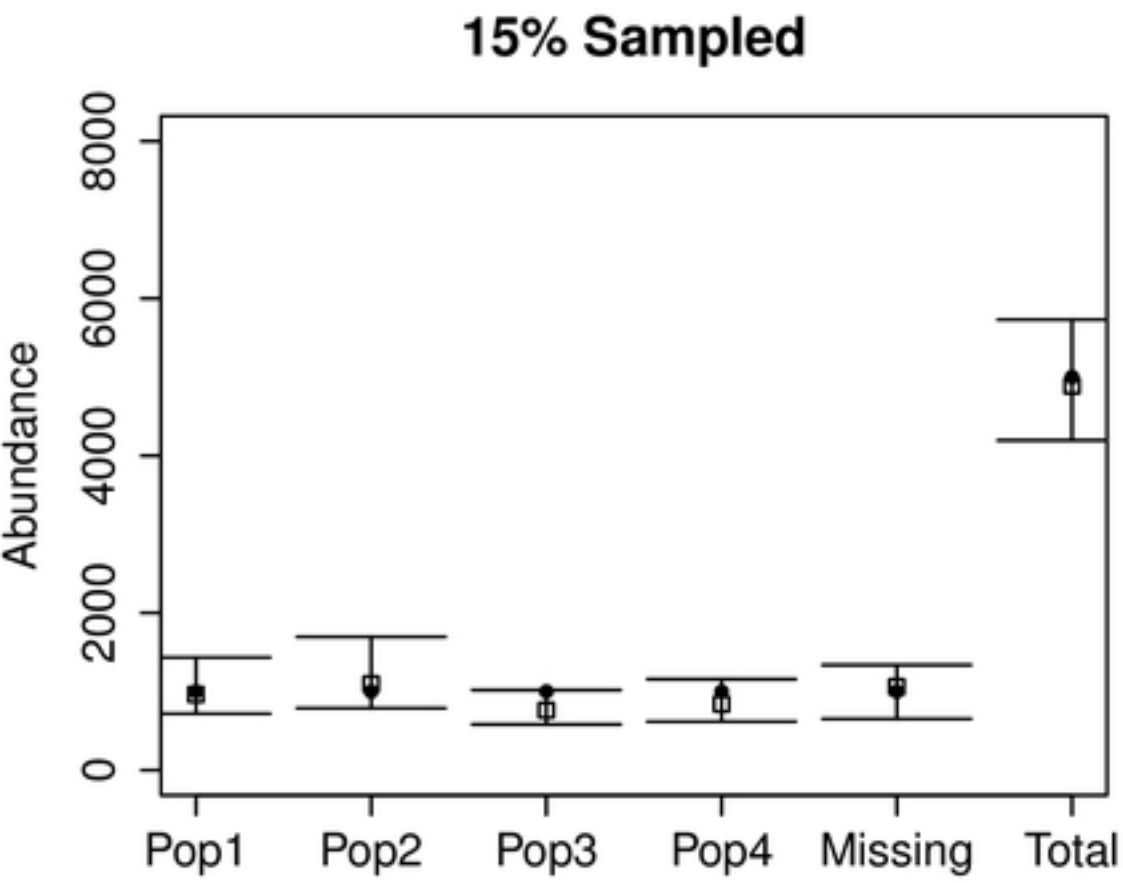
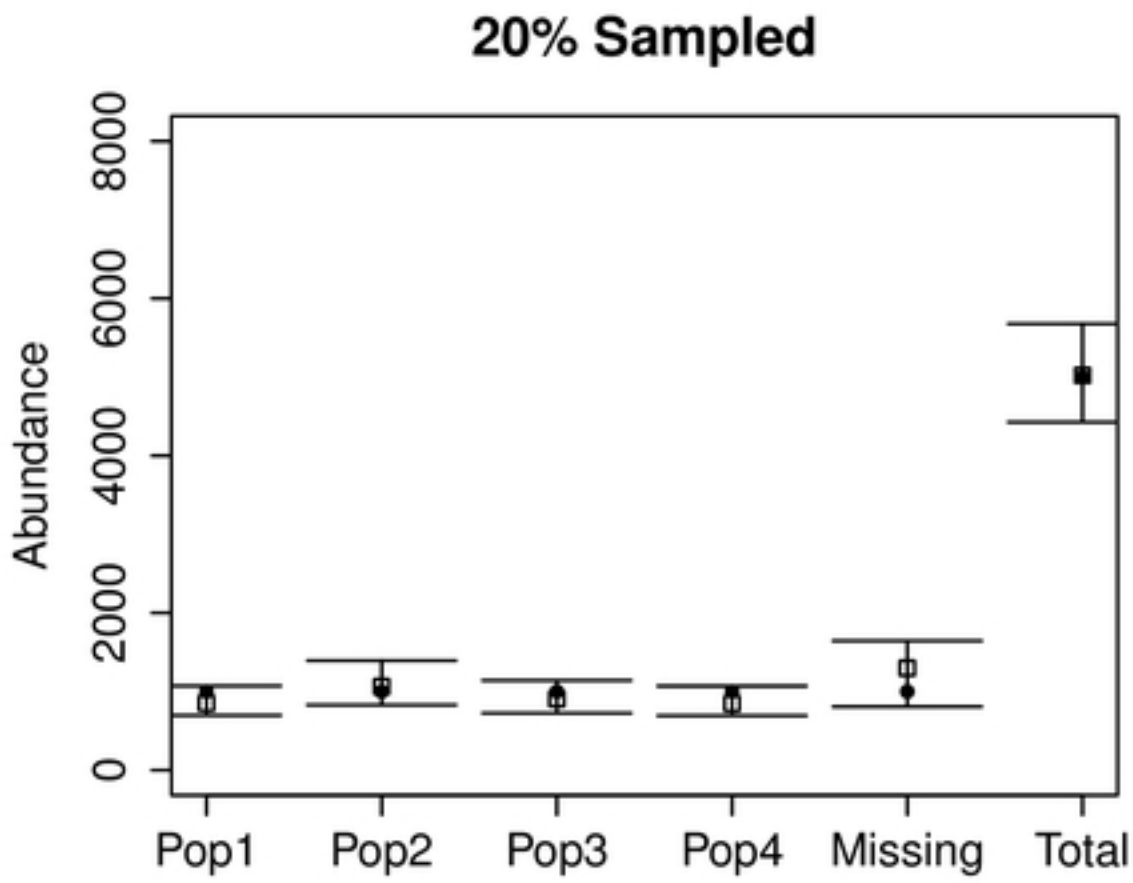


Figure 4


5% Sampled
Abundance
0
2000
4000
6000
8000
Pop1
Pop2
Pop3
Pop4
Missing
Total


10% Sampled
Abundance
0
2000
4000
6000
8000
Pop1
Pop2
Pop3
Pop4
Missing
Total


15% Sampled
Abundance
0
2000
4000
6000
8000
Pop1
Pop2
Pop3
Pop4
Missing
Total


20% Sampled
Abundance
0
2000
4000
6000
8000
Pop1
Pop2
Pop3
Pop4
Missing
Total

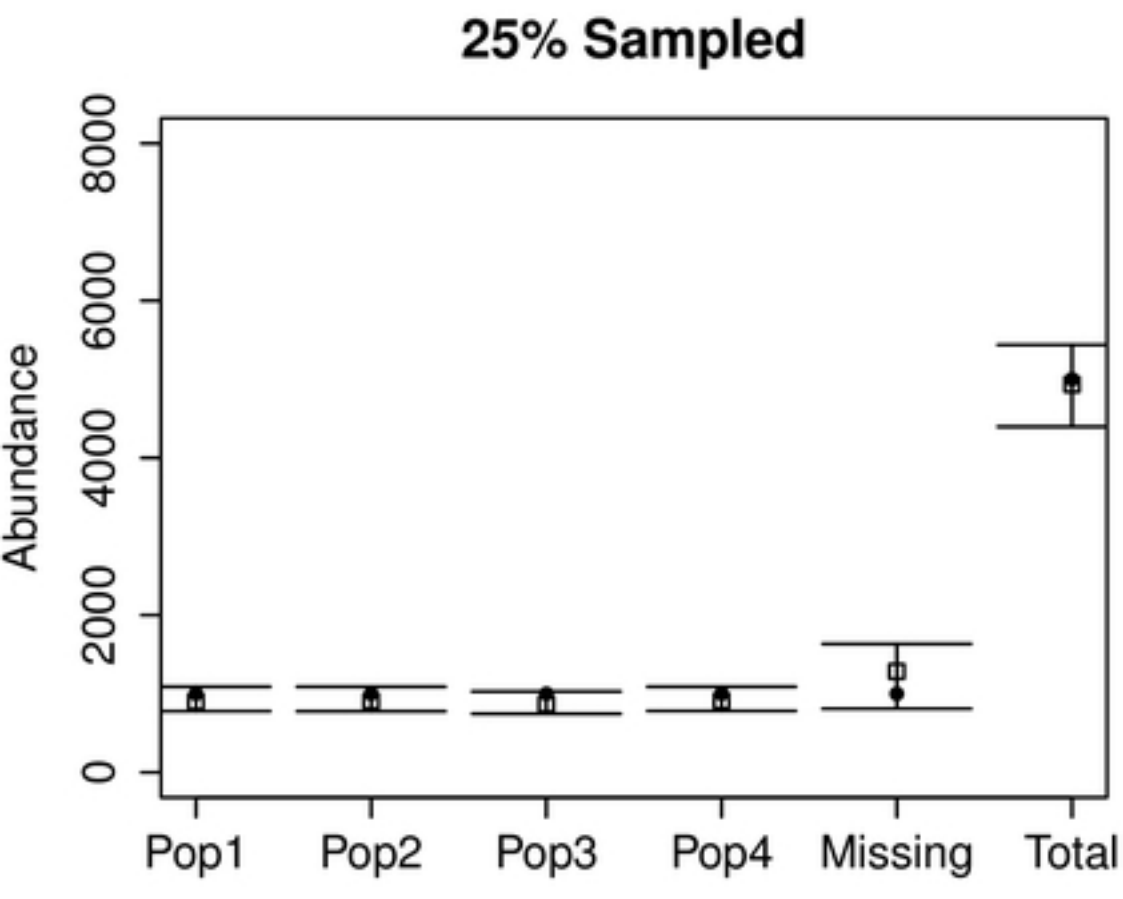

25% Sampled
Abundance
0
2000
4000
6000
8000
Pop1
Pop2
Pop3
Pop4
Missing
Total

Figure 5

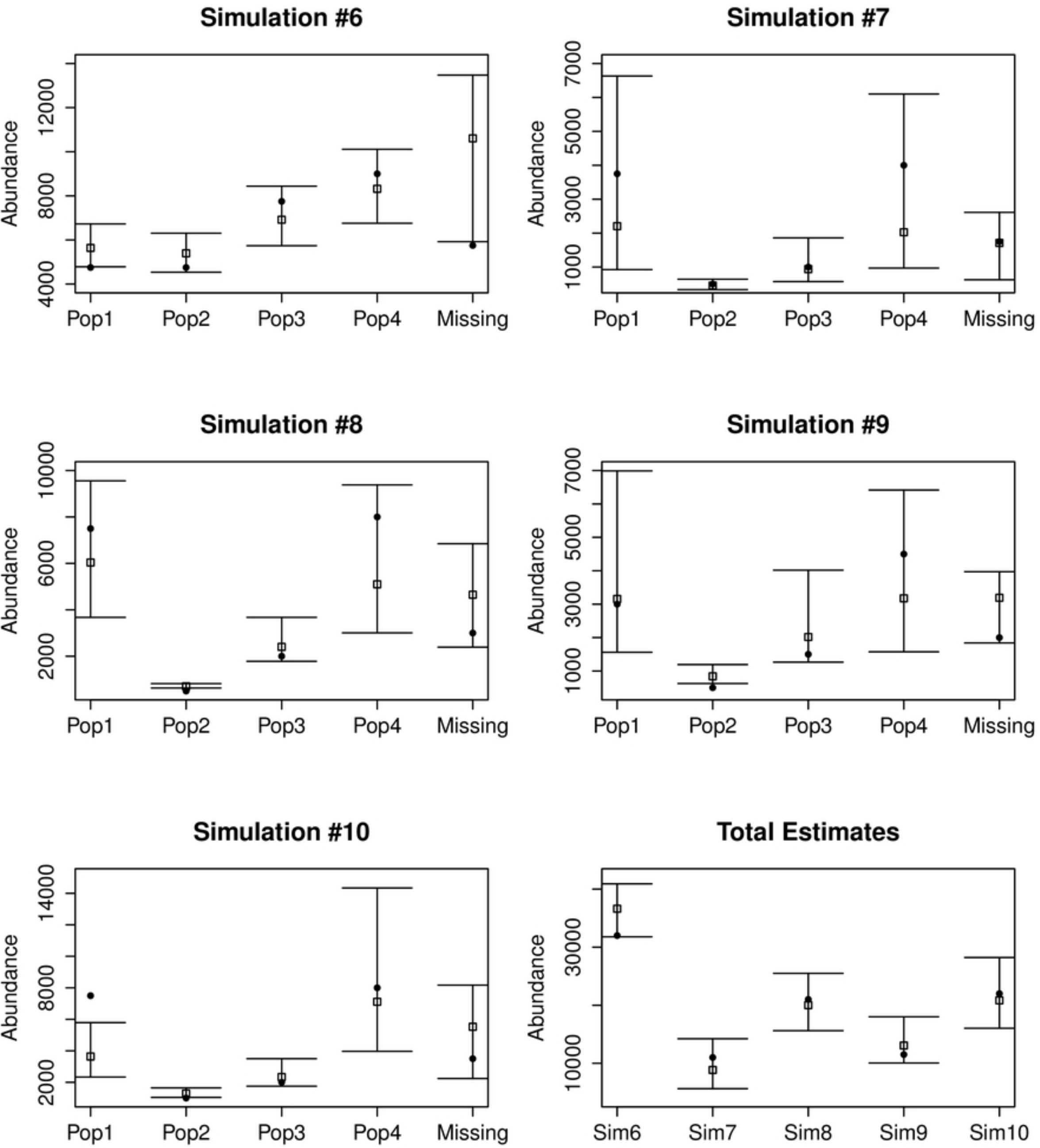

Simulation #6
Abundance
4000
8000
12000
Pop1
Pop2
Pop3
Pop4
Missing
Simulation #7
Abundance
1000
3000
5000
7000
Pop1
Pop2
Pop3
Pop4
Missing
Simulation #8
Abundance
2000
6000
10000
Pop1
Pop2
Pop3
Pop4
Missing
Simulation #9
Abundance
1000
3000
5000
7000
Pop1
Pop2
Pop3
Pop4
Missing
Simulation #10
Abundance
2000
8000
14000
Pop1
Pop2
Pop3
Pop4
Missing
Total Estimates
Abundance
10000
30000
Sim6
Sim7
Sim8
Sim9
Sim10

Figure 6