# File S4 Methodology for estimating the critical population size

We generated 1000 pseudo-observed data sets and 100,000 simulated data sets, each with independent introductions to 200 locations. The critical population sizes were drawn from a uniform distribution on [0,100]. We fixed the other parameters of the population dynamics ($k_0 = 10,000, k_1 = 1000, r = 0.1, c = 0$) and assumed them to be known with certainty. We further assumed that the original distribution of founder population sizes was Poisson with mean 20, and sampled the founder population sizes independently for each location from the conditioned distribution of founder population sizes for the respective critical population size. Given the selected founder population size, we simulated the population dynamics at each location from the conditioned Markov chain until the population reached size 200, i.e. twice the largest possible critical population size.

At this point, we sampled $n_l = 20$ individuals at both genome copies, resulting in 40 copies of each locus from a given location. We generated genealogies for 20 freely recombining loci. To obtain a more differentiated picture of patterns of genetic variation and capture as much information as possible, we did not use the average pairwise coalescence times or total lengths of the genealogy as before. Instead, we took the means and variances across loci of the entries of the site-frequency spectrum (SFS) $\xi_i$, i.e. the number of mutations that appear in $i$ chromosomes in the sample for $i \in 1, 2, \ldots, 39$. To compute these summary statistics, we first took the combined length of all branches $B_i$ that have $i$ descendants in the sample and assumed the infinite-sites model such that the number of mutations on a branch of length $b$ is Poisson-distributed with parameter $\mu \cdot b$. Given sufficiently many loci, which we assume we have, we do not actually need to simulate mutations along the branches. Instead, we can use the branch lengths to directly estimate the means and variances across loci of the $\xi_i$ as

$$\hat{\mathbf{E}}[\xi_i] = \mu \cdot \bar{B}_i \tag{S8}$$

and, using the law of total variance,

$$\hat{\mathbf{Var}}[\xi_i] = \hat{\mathbf{E}}\left[\mathbf{Var}[\xi_i|B_i]\right] + \hat{\mathbf{Var}}\left[\mathbf{E}[\xi_i|B_i]\right] = \mu \cdot \bar{B}_i + \mu^2 \cdot s^2(B_i), \tag{S9}$$

where the $\bar{B}_i$ are the average branch lengths across the $n_l$ loci and the $s^2(B_i)$ are the corresponding empirical variances. We assumed throughout that $\mu = 0.001$.

We further summarized the data for each SFS entry $i \in 1, 2, \ldots, 39$ by computing the averages and empirical standard deviations of the quantities in (S8) and (S9) across locations. To investigate how the quality of the estimation depends on the number of independent locations available, we took into account either only 10, 25, 50, 100, or all 200 of them to compute these statistics. Using the pls script from abctoolbox (Wegmann *et al.* 2010) and the pls package in R (Mevik and Wehrens 2007), we then conducted partial least squares regression on the first 10,000 simulated data sets to condense

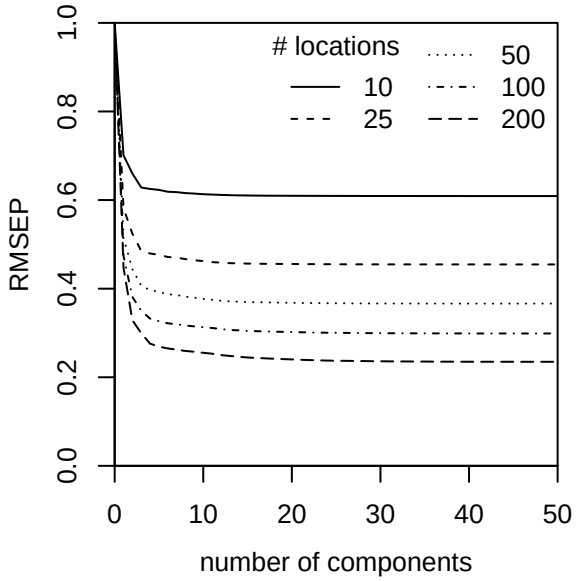


**Figure S5** Root mean squared error of prediction (RMSEP) as a function of the number of PLS components for various numbers of locations.

the information contained in the 156 summary statistics to a smaller number of components. To decide on the number of components, we examined plots of the root mean squared error of prediction (RMSEP) as a function of the number of components (Figure S5). For none of the different numbers of locations did the RMSEP change substantially beyond 20 components. Thus, we decided to include 20 components as summary statistics for ABC.

We used these 20 PLS components as summary statistics for parameter estimation with the R package abc (Csilléry *et al.* 2012). We chose a tolerance of 1 % and used the option "loclinear" implementing the local linear regression method (Beaumont *et al.* 2002). To avoid estimated parameter values that fall outside the prior, we estimated $\ln(a/(100-a))$ and then back-transformed the estimated values. For each pseudo-observed data set, we thus used the 100,000 simulated data sets to approximate the posterior distribution of the critical population size given a uniform prior on [0,100]. For each data set, we stored the mean of the posterior, which we take as our point estimator, and the 50 % and 95 % credibility intervals. We observed that the quality of parameter inference improved with an increasing number of locations (Figures 8 and S6). An examination of the percentage of pseudo-observed data sets for which the true parameter value falls into the respective 50 % or 95 % credibility interval suggests that ABC approximates Bayesian inference reasonably well in this case (Figure S7).

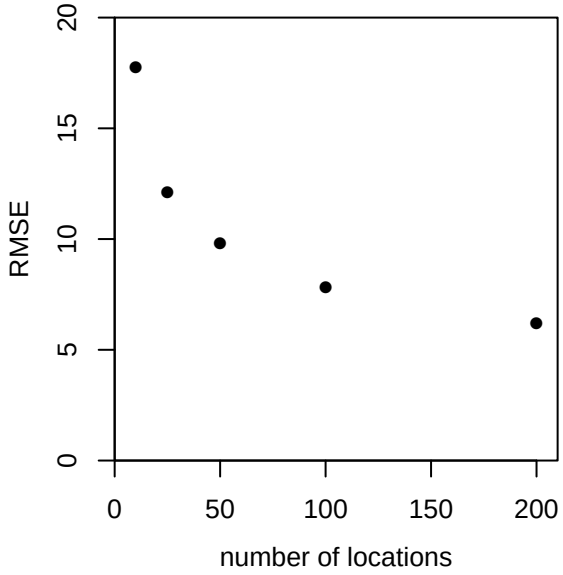


**Figure S6** The root mean squared error (RMSE) of the estimated critical population size as a function of the number of independent locations used for the estimation.

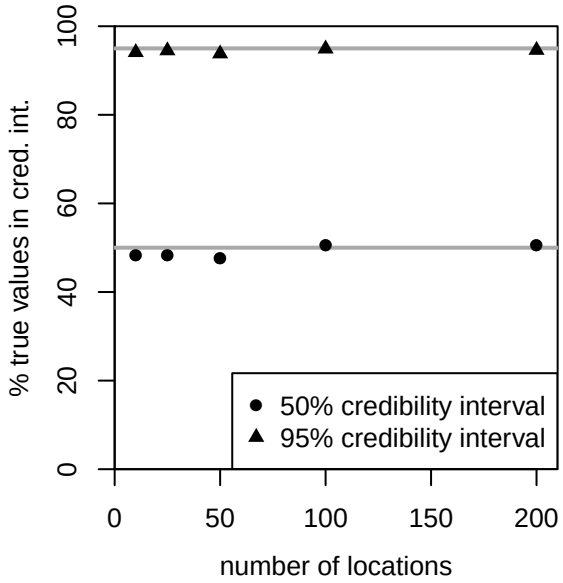


**Figure S7** Percentage of true parameter values that fall within the 50% and 95% credibility interval, an indicator for how well Approximate Bayesian Computation approximates Bayesian inference. The gray lines are at 50% and 95%.