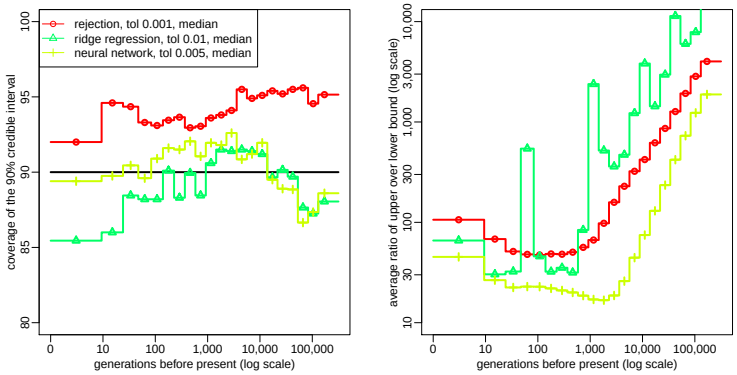
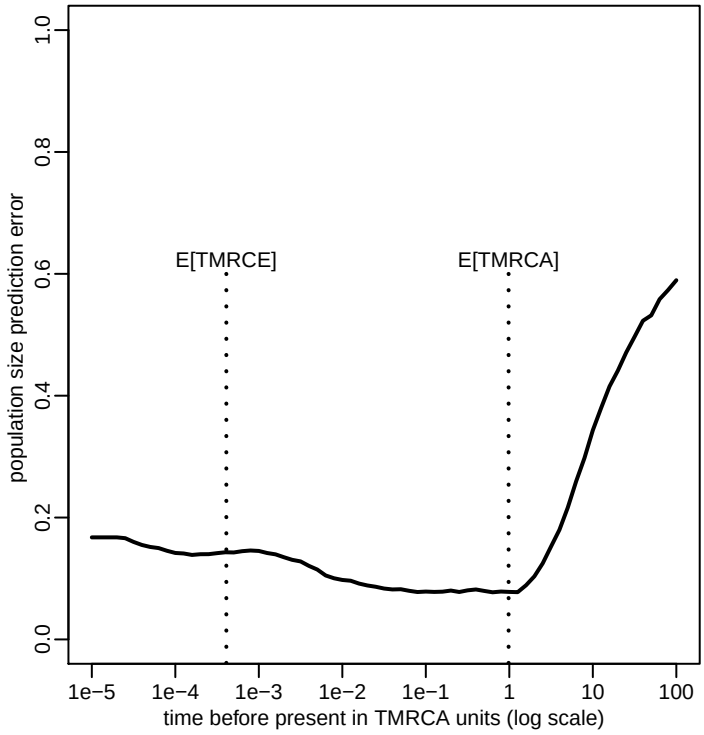
Inferring population size history from large samples of genome wide molecular data - an ABC approach : Supporting Information.

S1 Fig



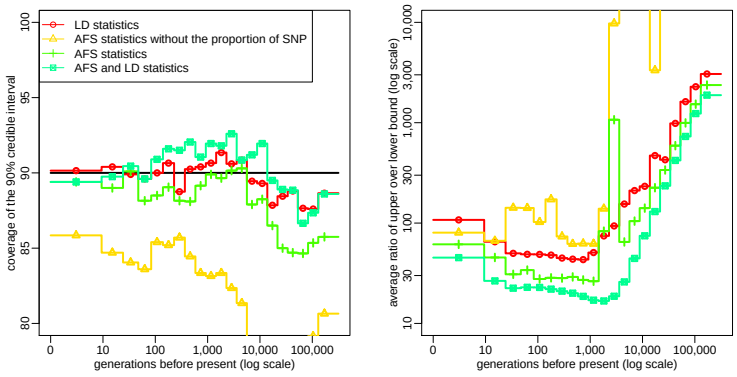
Accuracy of credible intervals obtained by ABC: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. The empirical coverage is the proportion of simulated histories for which the true population size was included in the 90% credible interval of the posterior distribution. If the posterior distribution was correctly estimated, this proportion should have been 90%, as shown by the black horizontal solid line. Parameter settings were the same as in Fig. 1.

S2 Fig



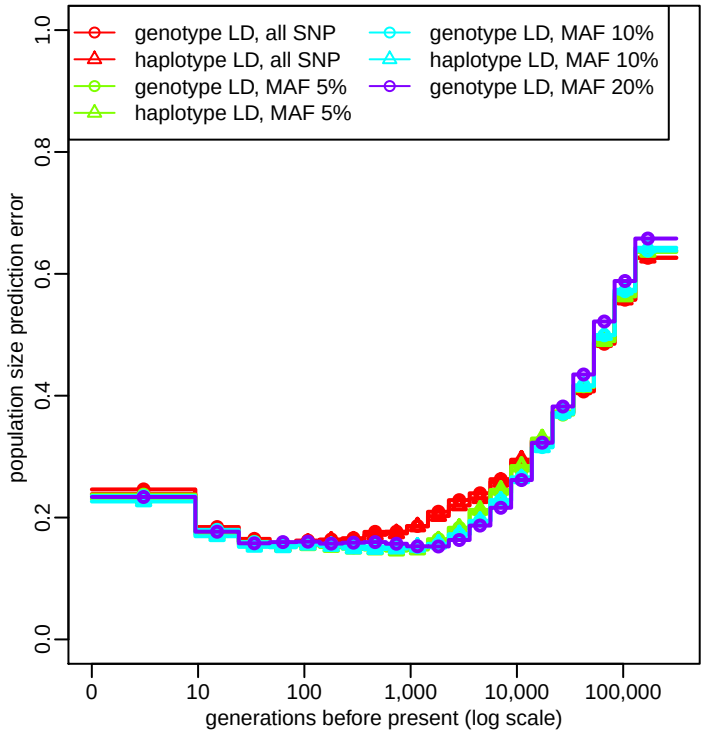
Accuracy of ABC estimation along the coalescent process: Prediction error for the estimated population size when time is measured in units of the expected time to the most recent common ancestor (TMRCA) of the sample. Prediction errors were evaluated from 2,000 random population size histories. Black vertical dotted lines indicate the expected time to the most recent coalescence event, $E[TMRCE]$, and the expected TMRCA, $E[TMRCA]$. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS statistics and SNPs with a MAF above 20% for LD statistics. The posterior distribution of each parameter was obtained by neural network regression [1], with a tolerance rate of 0.005. Population size point estimates correspond to the median of the posterior distribution.

S3 Fig



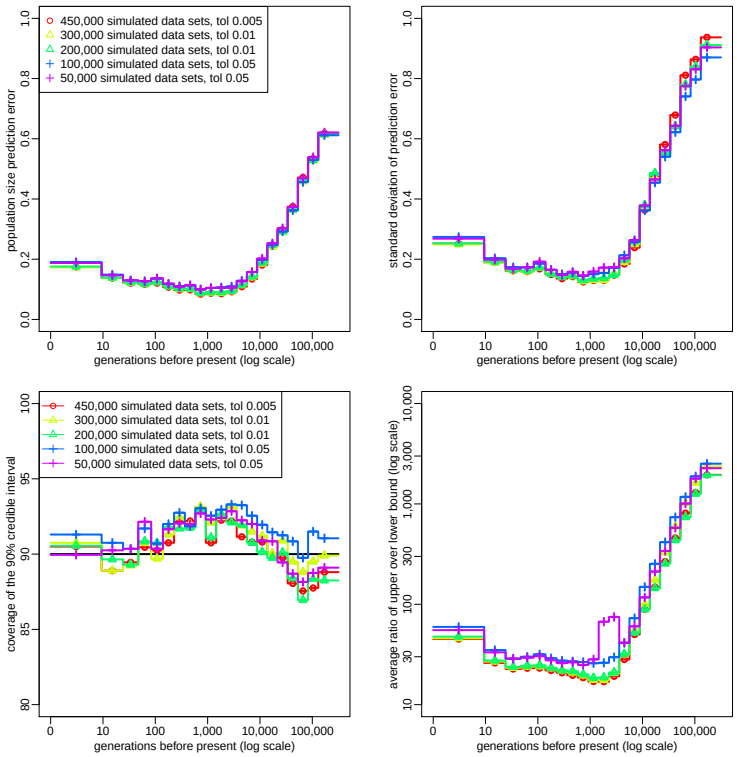
Accuracy of credible intervals obtained by ABC and relative importance of the summary statistics: Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. Parameter settings were the same as in Fig. 2. The very large credible intervals obtained on average with AFS statistics, in some time windows, are due to a retalively small number of PODs with extreme values.

S4 Fig



Accuracy of ABC estimation based on LD summary statistics: Prediction error for the estimated population size in each time window, evaluated from 2,000 random population size histories. Summary statistics considered in the ABC analysis were the average gametic LD (triangles) or the average zygotic LD (circles) for several distance bins. These statistics were computed from $n = 25$ diploid individuals, using different MAF thresholds. Other parameter settings were the same as in Figure 2.

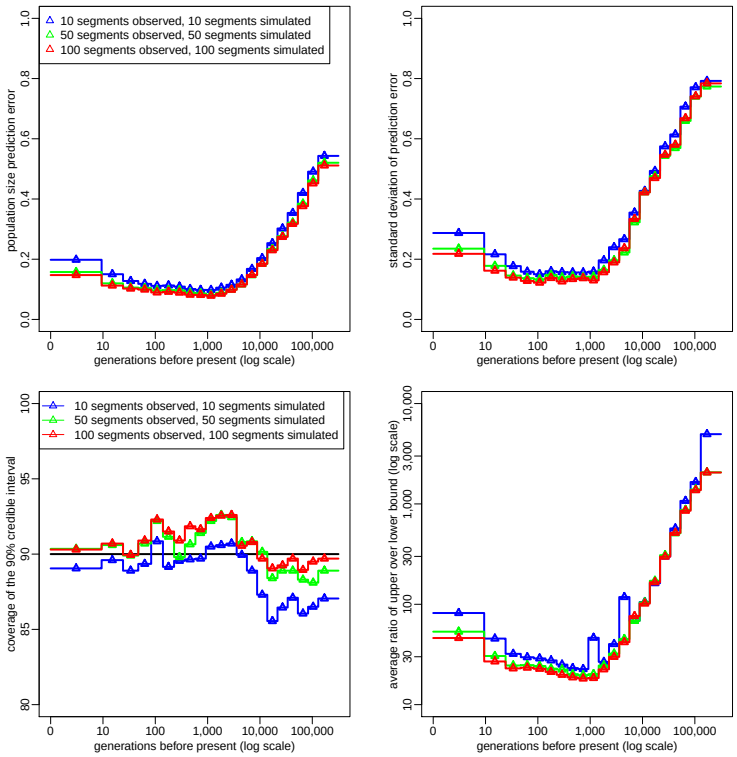
S5 Fig



Influence of the number of simulated data sets on ABC estimation: Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom : Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, for various numbers of simulated datasets (see the legend) with the same sample size ($n = 25$) and genome length (100 independent 2MB segments). Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. AFS statistics were computed using all SNPs and LD statistics were computed using SNPs with a MAF above 20%. The posterior distribution of each parameter was obtained by neural network regression, with the tolerance rate leading to the smallest prediction error. Population size point estimates

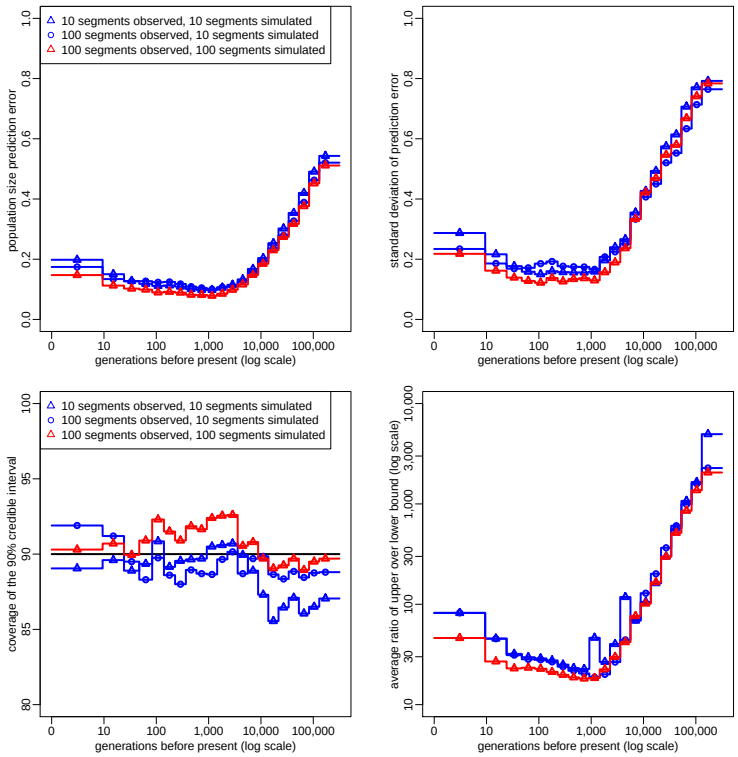
were obtained from the median of the posterior distribution.

S6 Fig



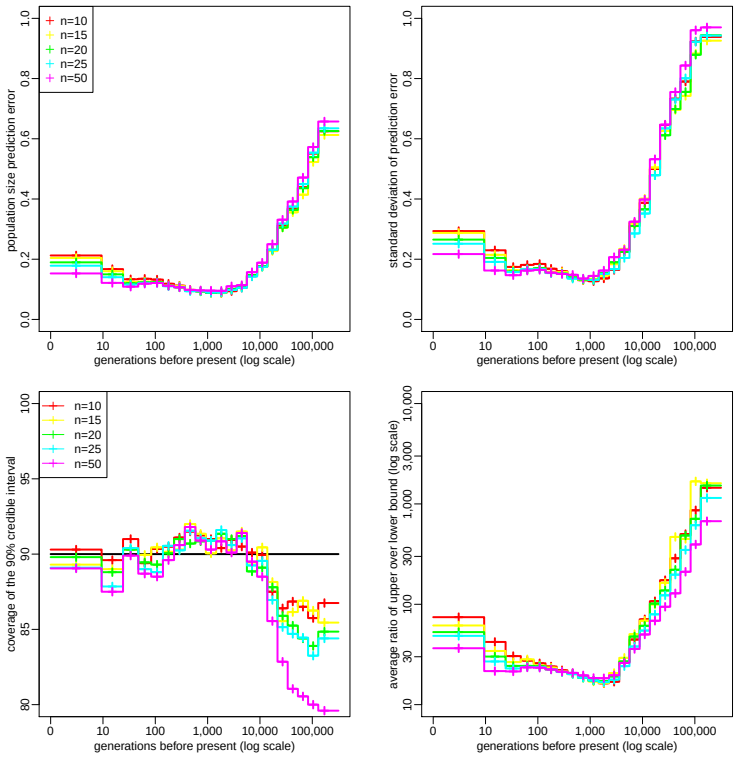
Influence of the genome length of simulated and observed data sets on ABC estimation: Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom : Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 10, 50 or 100 independent 2Mb-long segments (see the legend). Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size ($n = 25$) and genome length. The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. All other settings are similar to S5 Fig.

S7 Fig



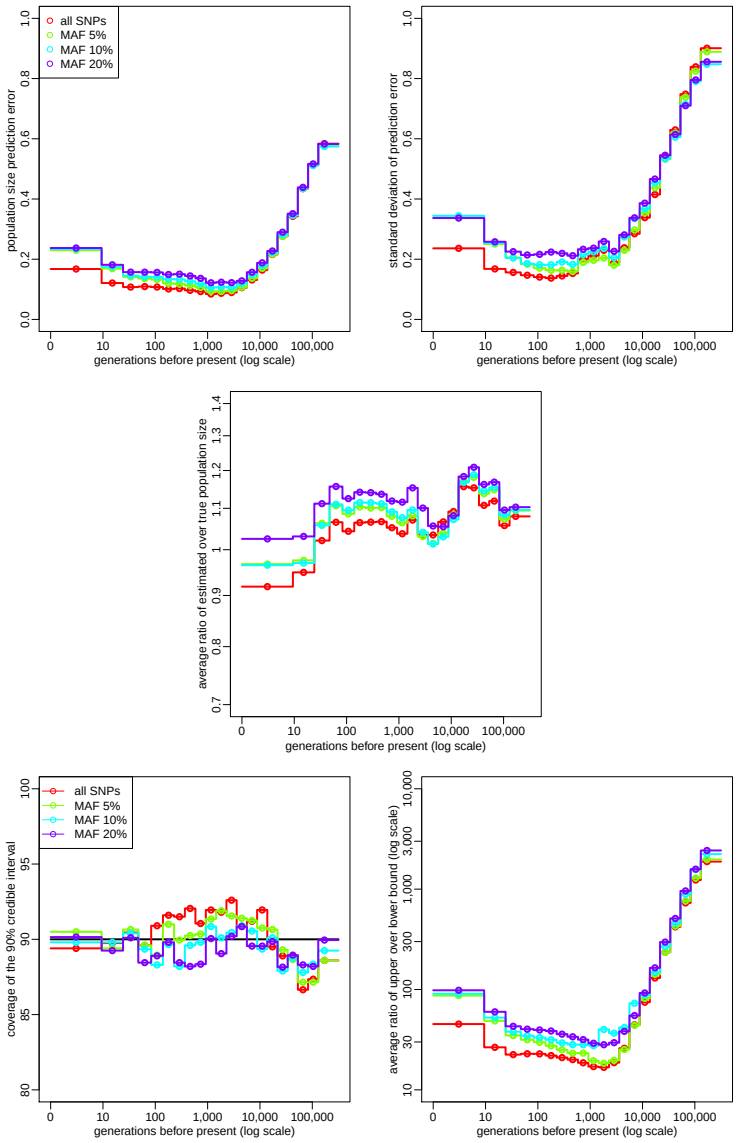
Using different genome lengths for simulated and observed data sets: Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom : Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 10 or 100 independent 2Mb-long segments (see the legend). Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size ($n = 25$) but a possibly different genome length (see the legend). The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. All other settings are similar to S5 Fig.

S8 Fig



Influence of the sample size on ABC estimation: Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Bottom : Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of n diploid genomes was simulated, for different values of n between 10 and 50 (see the legend). Each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size and genome length. All other settings are similar to S5 Fig.

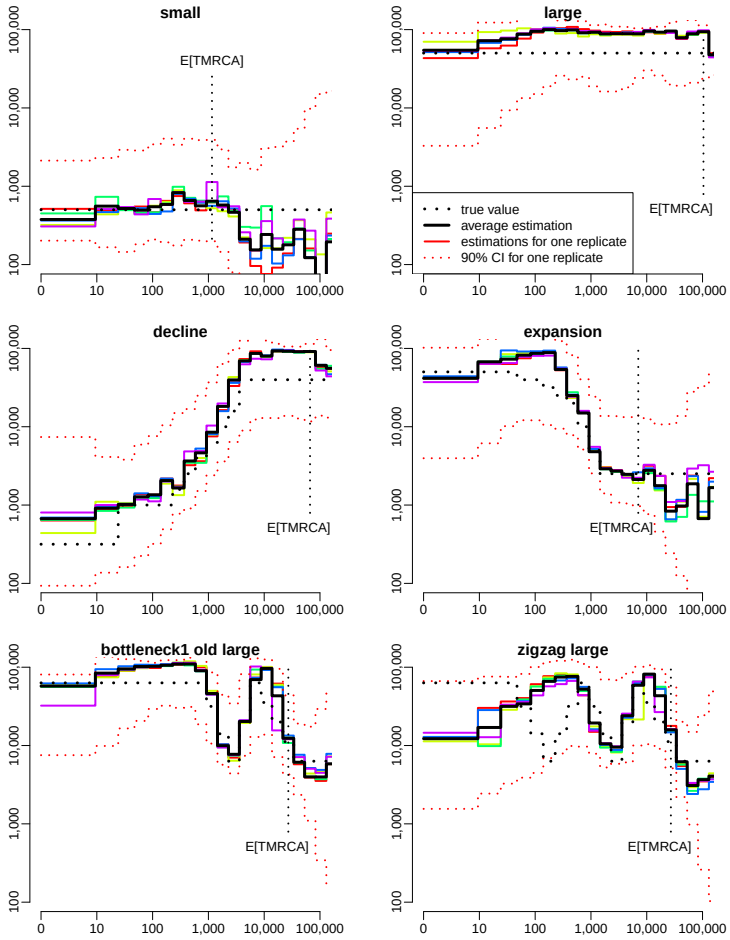
S9 Fig



Influence of MAF threshold on ABC estimation: Top: Prediction error for the estimated population size in each time window (left) and standard deviation of this error (right). Middle : Bias for the estimated population size in each time window. Bottom : Empirical coverage (left) and width (right) of the 90% credible interval for the population size in each time

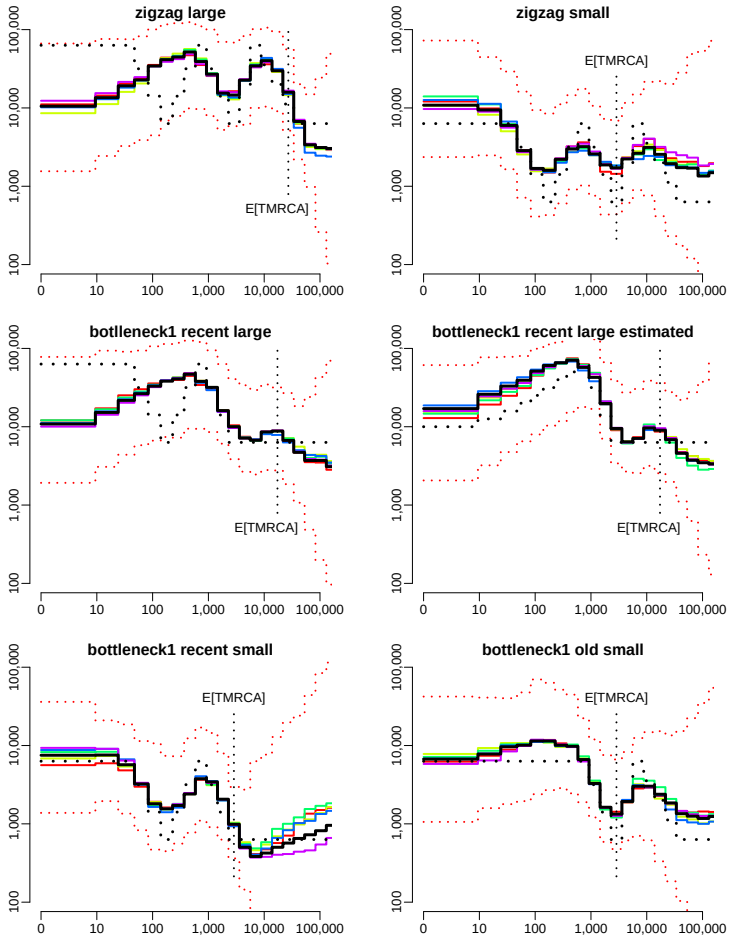
window. These quantiles were evaluated from 2,000 random population size histories. For each of these histories, one POD of $n = 25$ diploid genomes was simulated, where each genome consisted in 100 independent 2Mb-long segments. Population size history was estimated from this POD by ABC, using 450,000 simulated datasets with the same sample size and genome length. Summary statistics considered in the ABC analysis were (i) the AFS and (ii) the average zygotic LD for several distance bins. AFS statistics were computed using different MAF thresholds, LD statistics were computed from SNPs with a MAF above 20%. The posterior distribution of each parameter was obtained by neural network regression, with a tolerance rate of 0.005. Population size point estimates were obtained from the median of the posterior distribution.

S10 Fig



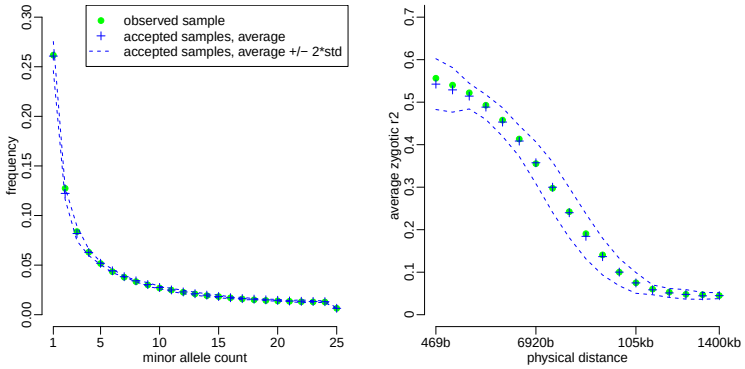
Estimation of population size history from the mode of the posterior distribution in six different simulated scenarios: All settings are similar to Figure 3, except that population size point estimates were obtained from the mode of the posterior distribution.

S11 Fig



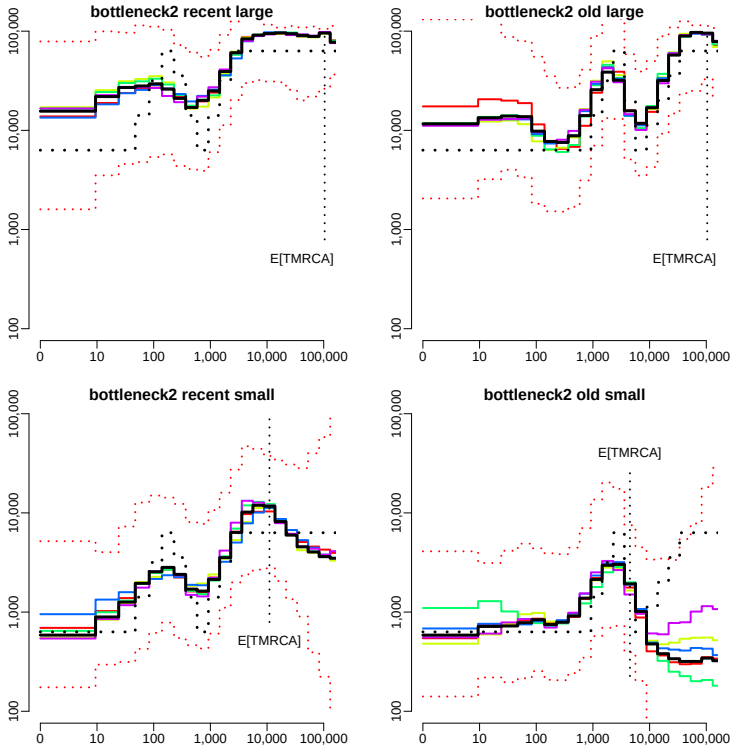
Estimation of population size history in the zigzag scenario and five related scenarios: a scenario where all population sizes are divided by ten compared to the original zigzag ("zigzag small", top right), a scenario where only the recent bottleneck of the original zigzag is simulated ("bottleneck1 recent large", middle left), a scenario corresponding to the history wrongly inferred by ABC based on data from the "bottleneck1 recent large" scenario (middle right), and two scenarios where only the recent (bottom left) or the old (bottom right) bottleneck of the "zigzag small" are simulated. All settings are similar to Fig. 3.

S12 Fig



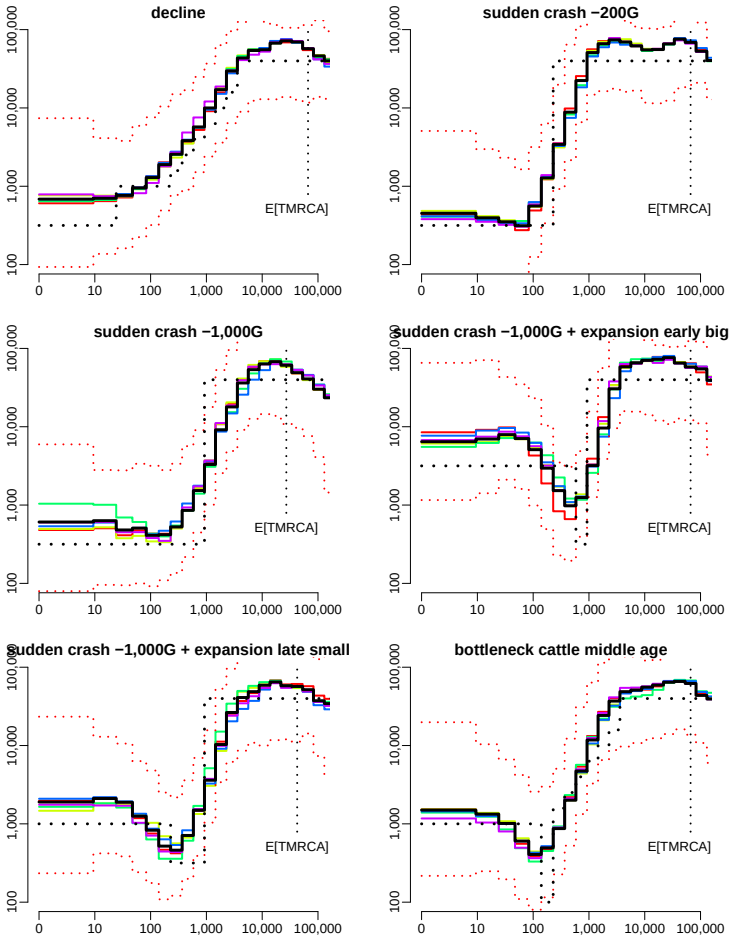
Observed and best simulated summary statistics in the "bottleneck1 recent large" scenario: For one of the five PODs analyzed in this scenario, observed AFS (left) and LD (right) statistics are shown by green full circles. The average value of these statistics over the five best simulated data sets, i.e. the five simulated data sets leading to the smallest distance between observed and simulated statistics, are shown by blue crosses. The variation of these statistics over the five best simulated data sets is also indicated by blue dotted lines, which correspond to the average value plus (or minus) twice the standard deviation of each statistic.

S13 Fig



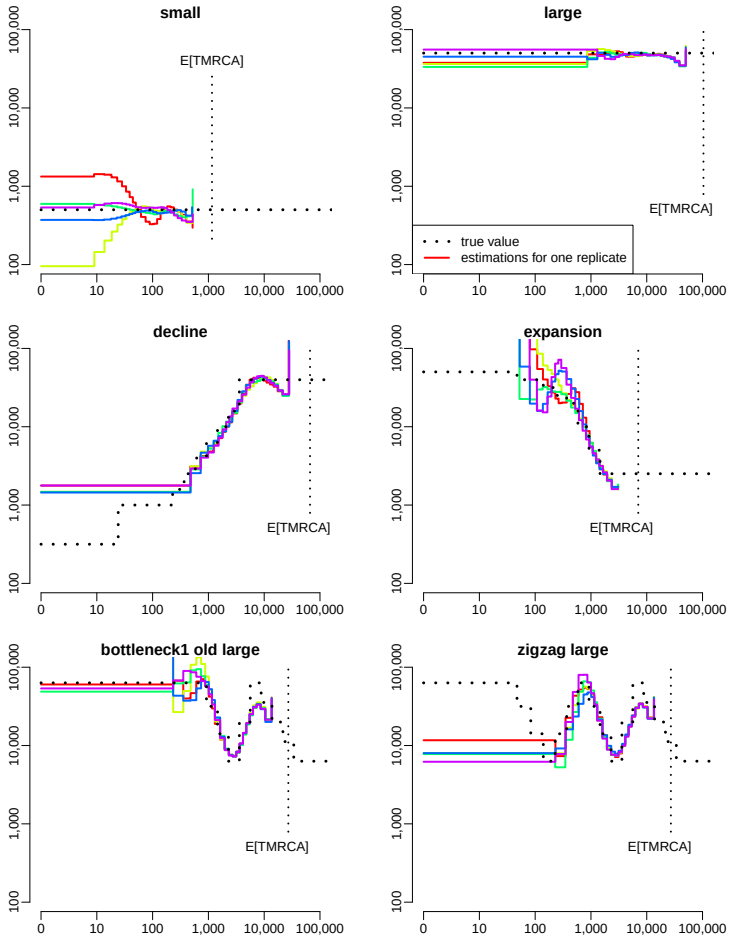
Estimation of population size history in four scenarios including a bottlenck followed by a population decline: Population size varied between 60,000 and 6,000 individuals in the top panels, and between 6,000 and 600 individuals in the bottom panels. Population size changes occurred between 2,300 and 50 generations BP in the left panels, and between 34,000 and 900 generations BP in the right panels. All settings are similar to Fig. 3.

S14 Fig



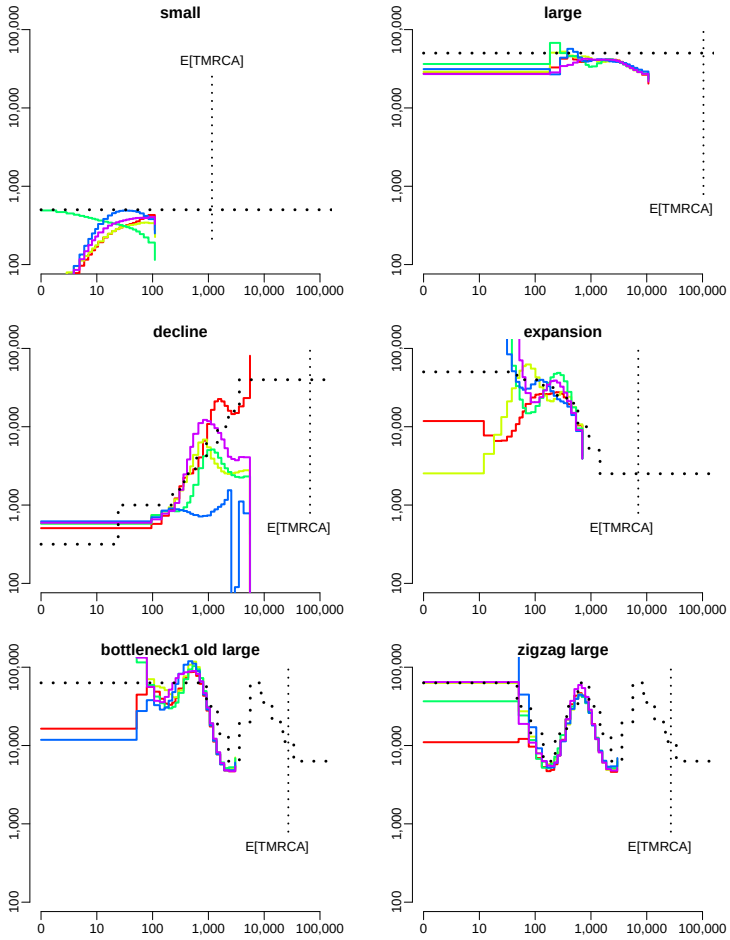
Estimation of population size history in the decline scenario and five related scenarios: a sudden (rather than continuous) decline from 40,000 to 300 individuals occurring 200 generations BP (top right), a sudden decline from 40,000 to 300 individuals occurring 1,000 generations BP (middle left), the same sudden decline followed by an expansion to 5,000 individuals occurring 580 generations BP (middle right) or an expansion to 1,000 individuals occurring 140 generations BP (bottom left), and a scenario similar to the continuous decline (top left) but including a sudden decline to 100 individuals between 230 and 140 generations BP, followed by an expansion to 1,000 individuals (bottom right). All settings are similar to Fig. 3.

S15 Fig



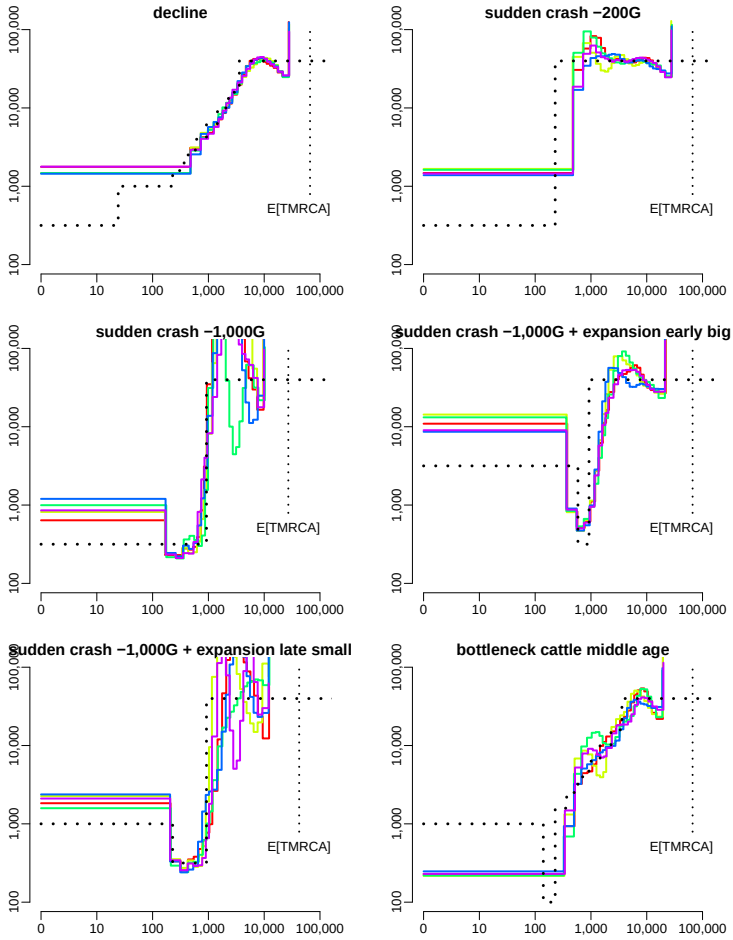
Estimation of past effective population size using MSMC with four haplotypes in six different simulated scenarios: For each scenario, the five PODs considered for MSMC estimation were the same as in Fig. 3. The expected TMRCA shown here is also the same as in Fig. 3, it corresponds to samples of 50 haploid sequences.

S16 Fig



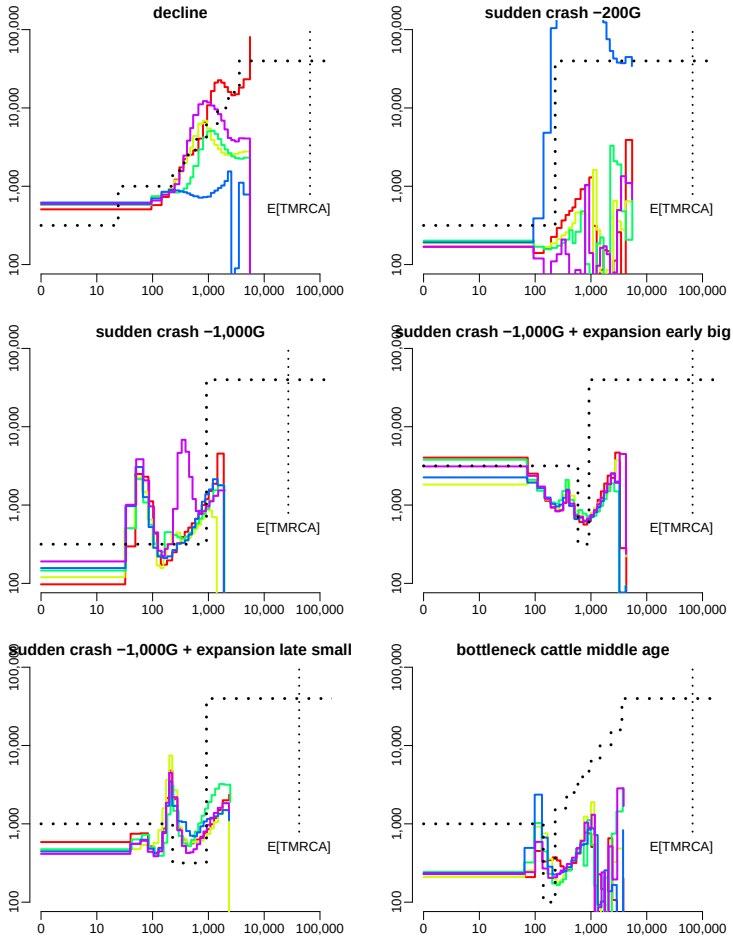
Estimation of past effective population size using MSMC with eight haplotypes in six different simulated scenarios: For each scenario, the five PODs considered for MSMC estimation were the same as in Fig. 3. The expected TMRCA shown here is also the same as in Fig. 3, it corresponds to samples of 50 haploid sequences.

S17 Fig



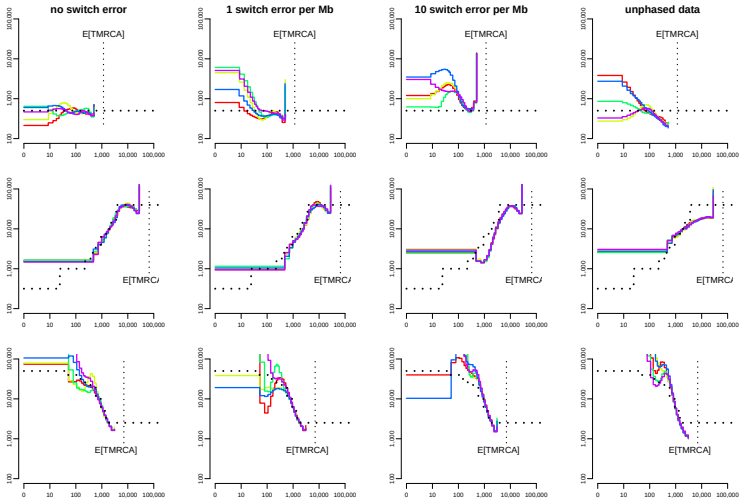
Estimation of past effective population size using MSMC with four haplotypes in the decline scenario and five related scenarios: For each scenario, the five PODs considered for MSMC estimation were the same as in S14 Fig. The expected TMRCA shown here is also the same as in S14 Fig, it corresponds to samples of 50 haploid sequences.

S18 Fig



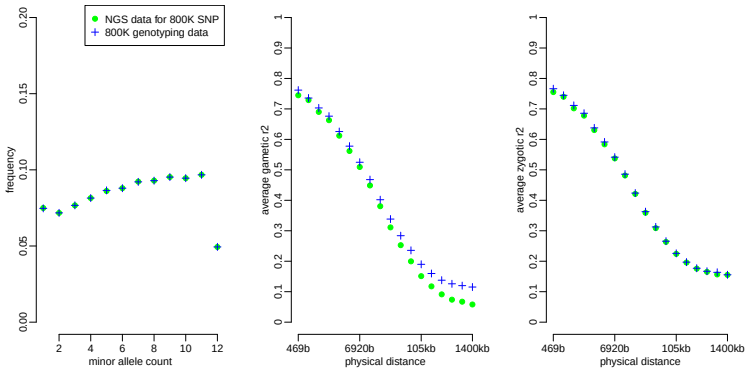
Estimation of past effective population size using MSMC with eight haplotypes in the decline scenario and five related scenarios: For each scenario, the five PODs considered for MSMC estimation were the same as in S14 Fig. The expected TMRCA shown here is also the same as in S14 Fig, it corresponds to samples of 50 haploid sequences.

S19 Fig



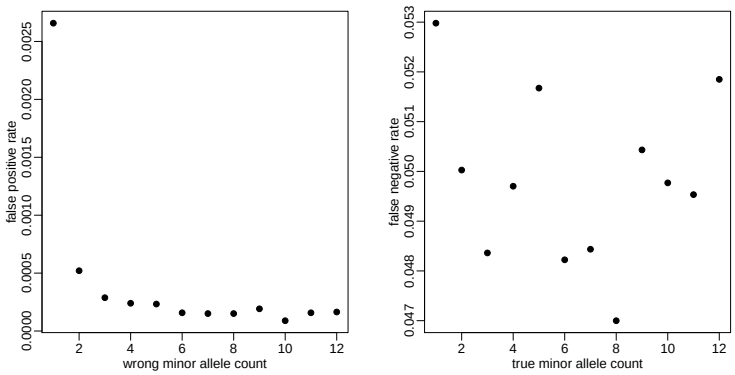
Influence of phasing errors on MSMC estimation: Estimation of past effective population size using MSMC with four haplotypes in the "small" scenario (top), the "decline" scenario (middle) and the "expansion" scenario (bottom). MSMC analyzes were run from perfectly phased data, phased data with 1 or 10 switch errors per Mb and diploid individual, or unphased data (i.e. two unphased diploid individuals). All other settings are similar to S15 Fig.

S20 Fig



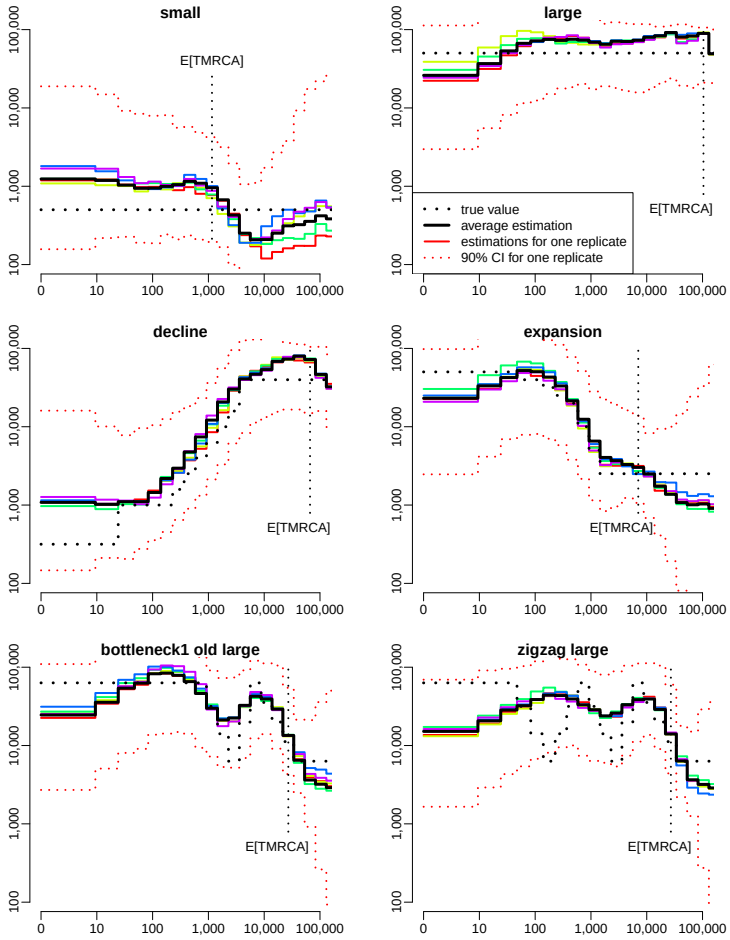
Comparison of summary statistics obtained from NGS and genotyping data: polymorphic site AFS, i.e. without the overall proportion of SNPs (left), average gametic LD (middle) and average zygotic LD (right). These statistics were computed from 12 Holstein animals for which both NGS data and genotyping data were available, using only SNP positions from the 800K chip (even for the NGS data statistics). No MAF threshold was used.

S21 Fig



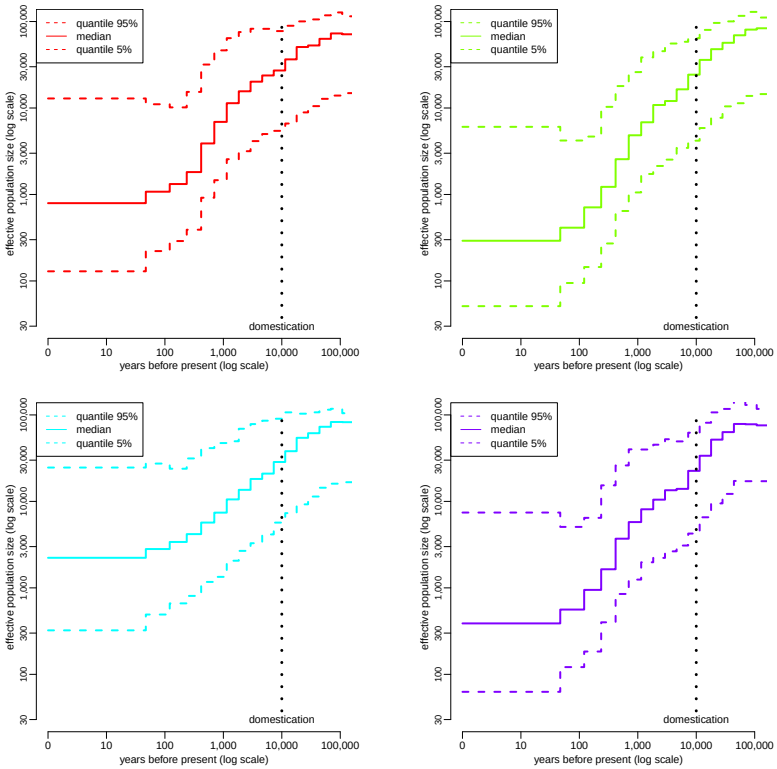
False positive and false negative rates of SNP detection in the 1,000 bull genome project: Error rates were computed from 12 Holstein animals for which both NGS data and genotyping data were available. False positive SNPs were positions that were found polymorphic in the NGS data but not in the 800K data. Their minor allele count in the NGS data was called the wrong minor allele count. False negative SNPs were positions that were found polymorphic in the 800K data but not in the NGS data. Their minor allele count in the 800K data was called the true minor allele count.

S22 Fig



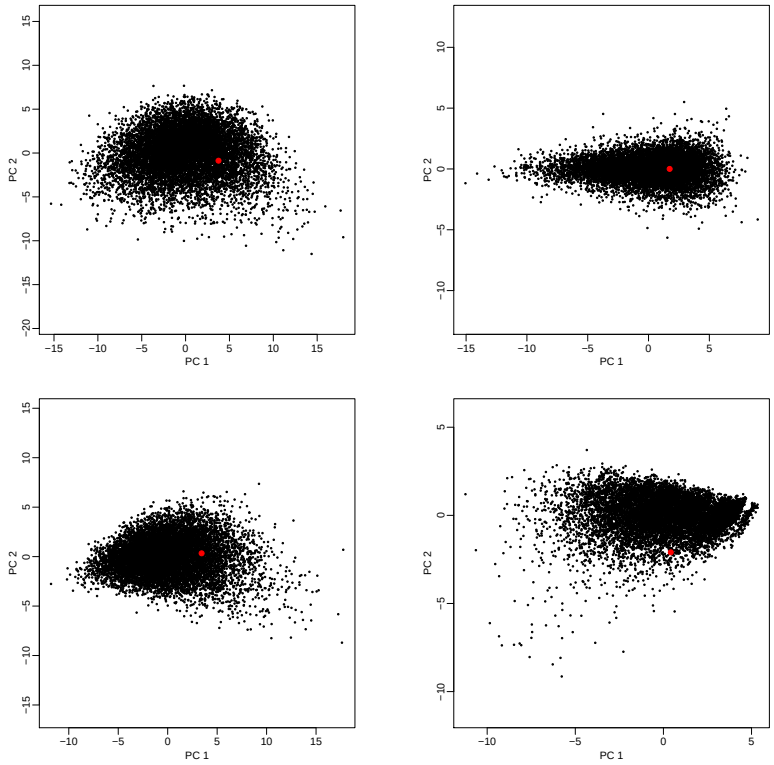
Estimation of population size history using ABC without rare SNPs in five different simulated scenarios: All settings are similar to Figure 3, except that AFS statistics were computed only from SNPs with a MAF above 20%.

S23 Fig



Ninety percent credible intervals of estimated population size history in four cattle breeds: Holstein (top left), Angus (top right), Fleckvieh (bottom left) and Jersey (bottom right). Parameter settings are the same as in Figure 6.

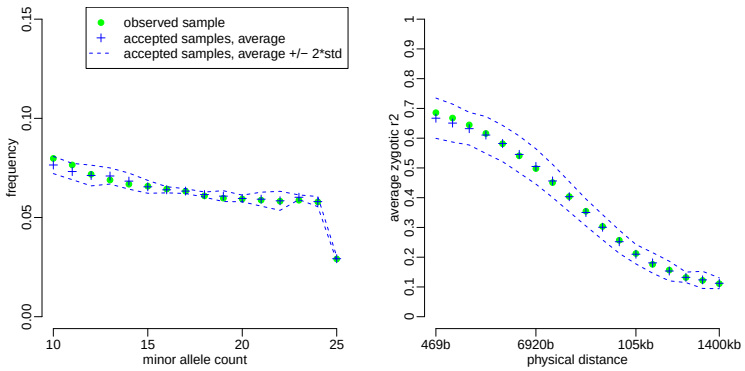
S24 Fig



Predictive posterior check of the population size history estimated in the Holstein cattle breed (Figure 6): Ten thousand genomic samples were simulated under population size histories that were sampled from the posterior distribution estimated in Figure 6. Four combinations of summary statistics were computed from each sample: AFS and LD statistics (top left), AFS statistics alone (top right), LD statistics alone (bottom left) and IBS statistics (bottom right, see the Methods for a detailed description of these statistics). For each of these combinations, a principal component analysis (PCA) of the 10,000 simulated samples was performed: the projection of all samples on the two first dimensions of this PCA are plotted in black. The vector of summary statistics observed in Holstein was then projected on the same hyperplan. It always fell within the cloud of simulated summary statistics, which shows that the estimated history is able to reproduce summary statistics that are indeed similar to the observed ones. Interestingly, this also holds for IBS statistics, which were not used for the estimation. Results are shown for the Holstein breed but they were similar for the other

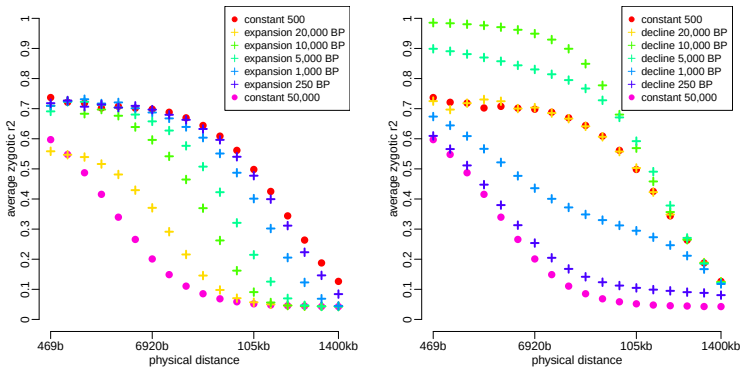
breeds.

S25 Fig



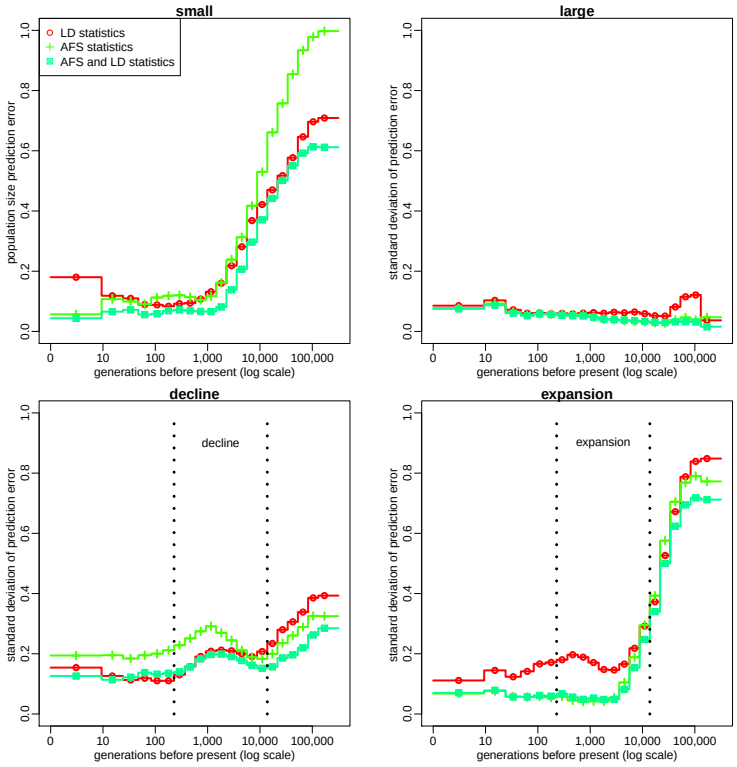
Observed and best simulated summary statistics in the Holstein cattle breed: Observed AFS (left) and LD (right) statistics are shown by green full circles. The average value of these statistics over the five best simulated data sets, i.e. the five simulated data sets leading to the smallest distance between observed and simulated statistics, are shown by blue crosses. The variation of these statistics over the five best simulated data sets is also indicated by blue dotted lines, which correspond to the average value plus (or minus) twice the standard deviation of each statistic.

S26 Fig



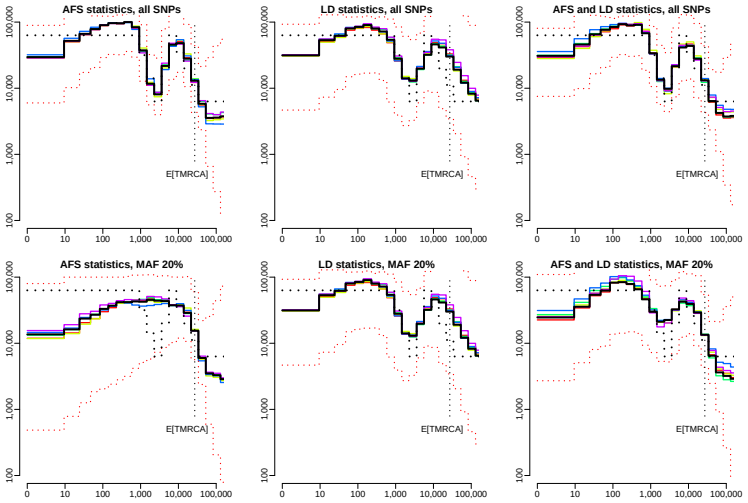
Influence of population size changes on LD statistics: LD statistics for several scenarios inplying a sudden expansion from 500 to 50,000 individuals (left) or a sudden decline from 50,000 to 500 individuals (right). Several expansion or decline times were considered, as well as two scenarios with a constant population size of 500 or 50,000 individuals (see the legend). For each scenario, LD statistics were averaged over 20 PODs including 25 diploid genomes and 100 2Mb-long regions. In contrast with expansion scenarios, some decline scenarios lead to even larger LD statistics than those obtained for a constant small population. Indeed, as these declines are very old compared to the expected TMRCA of a population of 500 individuals, their main effect is to increase, at some loci, the time during which the sample has only two ancestral lineages. Because this increase is very large (backward in time, population size, and thus expected coalescence time, are suddenly multiplied by 100), mutations occuring in this part of the coalescence tree eventually represent a large proportion of all oberved polymorphic sites. Besides, for two linked loci with similar topologies of the coalescence tree, mutations occuring in this part of the tree lead to very high r^2 values, up to 1 if the topologies are exactly the same.

S27 Fig



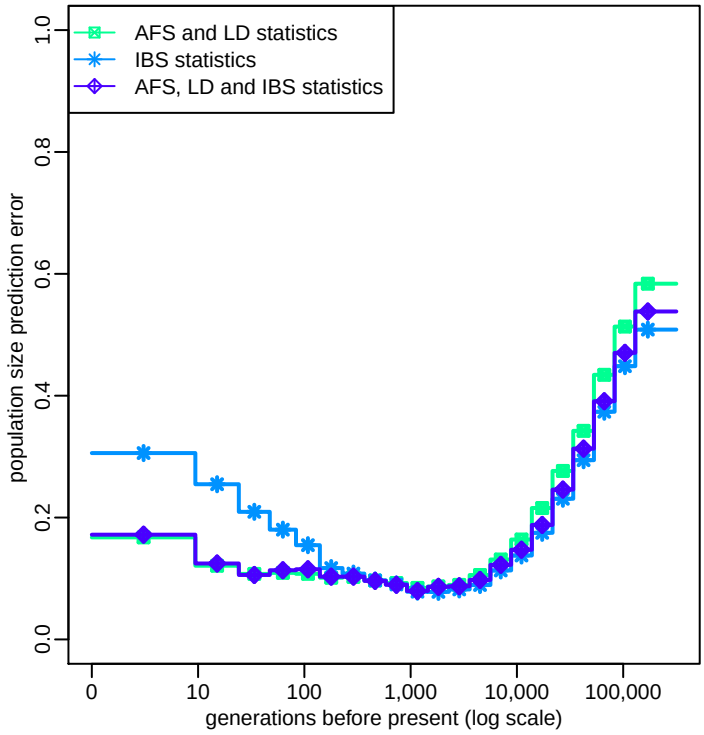
Accuracy of ABC and relative importance of LD and AFS in different families of scenarios: Prediction error for the estimated population size in each time window, focusing on scenarios with a population size below 1,000 (top left), above 10,000 (top right), below 1,000 in the last 200 generations and above 10,000 for times more ancient than 13,000 generations BP (bottom left) or above 10,000 in the last 200 generations and below 1,000 for times more ancient than 13,000 generations BP (bottom left). For the two latter scenarios, the time window where population size goes from above 10,000 to below 1,000 (or vice versa) is delimited by vertical dotted lines. For each scenario category, PE were evaluated from 2,000 random histories. Summary statistics considered in the ABC analysis were either the AFS statistics alone, the LD statistics alone or the AFS and LD statistics together (see the legend). All other settings are similar to Fig. 2.

S28 Fig



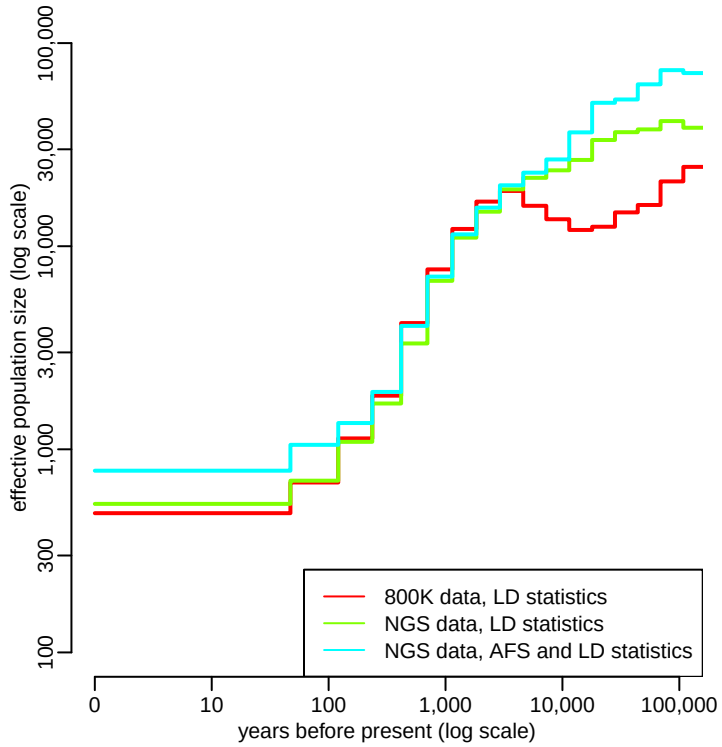
Estimation of population size history using different ABC settings in the "bottleneck1 old large" scenario: Summary statistics considered in the ABC analysis were either the AFS statistics alone (left column), the LD statistics alone (middle column), or the AFS and LD statistics together (right column). AFS statistics were computed using either all SNPs (top panels) or only those with a MAF above 20% (bottom panels). All other settings are similar to Fig. 3.

S29 Fig



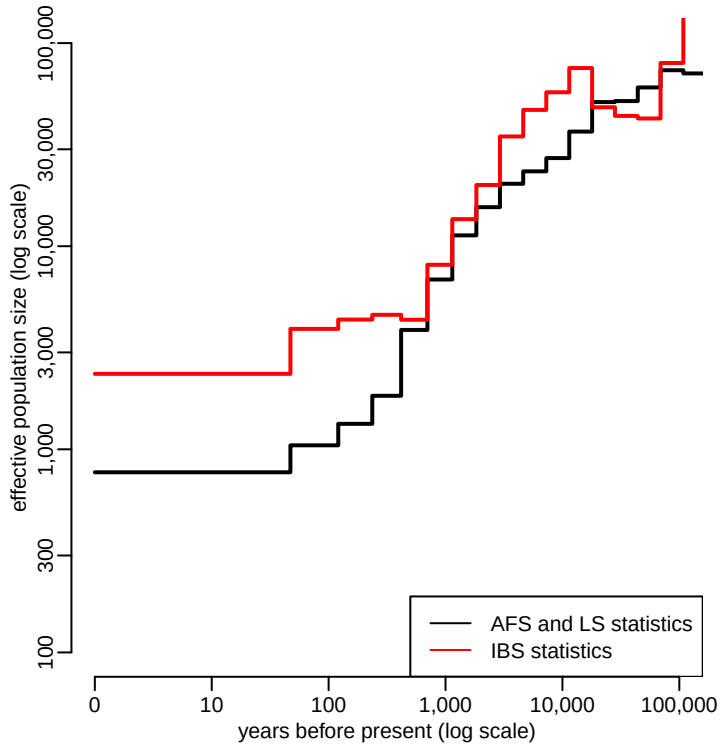
Accuracy of ABC estimation based on the distribution of IBS segment lengths: Prediction error for the population size in each time window, evaluated from 2,000 random population size histories. Summary statistics considered in the ABC analysis included several combinations of (i) the AFS, (ii) the average zygotic LD for several distance bins and (iii) the distribution of IBS segment lengths within one diploid individual. These statistics were computed from $n = 25$ diploid individuals, using all SNPs for AFS and IBS statistics and SNPs with a MAF above 20% for LD statistics. Other parameter settings are the same as in Figure 2.

S30 Fig



Added value of NGS for population size history estimation: Estimation of population size history in the Holstein cattle breed using ABC, based on whole genome NGS data from $n = 25$ animals. Summary statistics considered in the ABC analysis included different combinations of (i) the AFS and (ii) the average zygotic LD for several distance bins. These statistics were computed either from the SNPs that are included in the 800K SNP chip or from all SNPs found in the NGS data. A MAF threshold of 20% was used for all curves and statistics. Other parameter settings are the same as in Figure 5.

S31 Fig



Population size history in Holstein using IBS statistics: Estimation of population size history in the Holstein cattle breed using ABC, based on whole genome NGS data from $n = 25$ animals. Summary statistics considered in the ABC analysis were either both the AFS and the average zygotic LD for several distance bins, or the distribution of IBS segment lengths within one diploid individual. These statistics were computed using SNPs with a MAF above 20%. Other parameter settings are the same as in Figure 5.

References

[1] Blum M, Françcois O. Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing. 2010;20(1):63–73. Available from: http://dx.doi.org/10.1007/s11222-009-9116-0.