Supplementary Information for

Bayesian Inference Under the Multispecies Coalescent with ancient DNA sequences

Anna A. Nagel, Tomáš Flouri, Ziheng Yang, and Bruce Rannala

Corresponding Author: Anna A. Nagel, Bruce Rannala.
E-mail: aanagel@ucdavis.edu, brannala@ucdavis.edu

This PDF file includes:

Supporting Information Text

1. Simulation method

The simulation method in BPP was modified to accommodate serial sampling. The dates are specified in units of expected number of substitutions and given in an input file. Simulation works similarly the standard MSC simulation with a few extra steps. When simulating the MSC without tip dates, times for the coalescent events are drawn from an exponential distribution with the rate determined by the number of lineages within a population. When a coalescent event occurs, two lineages are randomly chosen to coalesce and the number of lineages decreases by one. This continues until either there is only one sequence in the population or the time drawn is older than the population divergence time. In either case, the time is reset to be the population divergence time, the number of lineages from the two populations are combined and the simulation continues backward in time until the root population only has one lineage. With tip dates, the simulation starts with the youngest sample time, rather than at time zero. Every time a coalescent time is drawn, it must be checked if the time is older than either the population divergence time or next oldest sampling event. In the former case, the simulation proceeds in the same way as without tip dating. In the later case, the time is set to the next oldest sampling event to determine all of the lineages that the sampling event are added to the lineage count, and the simulation proceeds.

2. Bayesian Simulations

A. MCMC settings. Bayesian simulations were conducted with 3000 replicate datasets. The parameters are described in the main text. Each MCMC was sampled 400,000 times, sampling every 4 iterations with 80,000 iterations of burn-in.

B. Convergence. Two MCMCs were run for each dataset to check convergence. Convergence was checked by comparing posterior samples from the two MCMCs for each set of parameters. A two-sample t-test was used to compare the posterior means in the two chains.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{2}{n}}}$$

where

$$s_p = \sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{2}}$$

In the standard two-sample t-test, X_i are the sample means, $s^2_{X_i}$ are the unbiased estimators of the variance and n is the sample size. There are $2n - 2$ degrees of freedom. Since the samples were not independent, rather than using the total number of samples in the MCMC, $n = 10,000$ was used as the sample size. The test was performed on estimates of all of the θs and τs. If there was a significant difference between the samples for any variable, the run was considered to not have converged. Additionally, any pairs of MCMCs that had effective sample sizes lower than 200 for any the θs and τs were considered to not have converged. This resulted in 408 datasets with MCMCs that did not converge. All runs that did not converge were re-run with different seeds and a burnin of 200,000 iterations. Convergence was checked again using the same criteria. There were 213 datasets that did not converge on this second analysis and these were excluded from the results (e.g. the plot summaries in Fig. S2, S3).

Assessing convergence of MCMC is non-trivial, and these methods of checking convergence were spot checked for MCMCs that did or did not converge. The trace plot, the effective sample sizes, and plots of kernel density estimation were further visually examined for these spot checked cases.

3. Simulations

A. MCMC settings. All MCMCs were sampled 400,000 of times, sampling every 4 of iterations. The burnin was 160,000 iterations. Two independent MCMCs were run for each dataset. Convergence was checked comparing the results between the independent MCMCs. See Materials and Methods for details of simulations parameters.

B. Convergence. Convergence was checked using criteria similar to the Bayesian simulations, except that an n of 2000 was used in the two-sample t-test and differences in the means of all parameters (θs, τs, $\tau^{\triangle}$s, and μ) were required to not be significantly different between the replicate MCMCs. All parameters except μ required an effective sample size of at least 200 in both MCMC replicates to be considered as converged. Runs that did not converge were re-run with different seeds and 600,000 samples, sampled every 4th iteration. The burnin length was not changed. The same test was conducted after re-running the MCMCs, except that the ancestral population sizes and the root age in expected number of substitutions were not checked and a two-sample t-test sample size of $n = 200$ was used. These parameters converged more slowly than other parameters, and were not central to the results. The root age in time before present was included in the convergence criteria and appeared to converge more quickly than root age in expected number of substitutions in some cases. The mitochondrial simulations and the recent population divergence simulations that did not meet the convergence criteria were removed from the results. These comprised no more than half of any set of 20 replicate simulations. The other MCMCs that did not meet the convergence criteria were re-run with different seeds and 1,200,000 samples, sampled every 4th iteration. These tended to be the larger

datasets with 500 or 2000 loci. The convergence was assessed again with the same test that was used for the first MCMC re-runs (ancestral population sizes and the root age in expected number of substitutions were not checked and a two-sample t-test sample size of $n = 200$ was used). The simulations that did not meet the convergence criteria were removed from the results. These simulations comprised no more than half of simulation replicates for any set of simulation parameters.

4. Empirical Analysis

A. Priors for nuclear dataset. To choose appropriate parameters for the root age prior, all of the loci were concatenated for each species. The average pairwise divergence between sequences from the mammoth and elephant species and the mastodon was calculated to specify a prior for the dataset with the mastodon. The average pairwise divergence between all pairs of species that are not sisters was calculated to choose a prior for the dataset without the mastodon. Gaps were removed from the two sequences being compared prior to calculating pairwise divergence and "n" was treated as a gap. When ambiguity codes existed in the sequences, equal probability was given to all possible bases indicated in the ambiguity code. This method will give an overestimate of root age, as the coalescent times must be older than the speciation time. However, this should give a reasonable order of magnitude for the prior mean. The variance was chosen such that there was a broad distribution around the mean, since there is not strong prior information about the speciation times in expected number of substitutions.

Species 1	Species 2	pairwise divergence
Asian	Forest	0.0072
Asian	Savannah	0.0070
Mammoth	Forest	0.0069
Mammoth	Savannah	0.0068
Asian	Mastodon	0.037
Forest	Mastodon	0.036
Mammoth	Mastodon	0.036
Savannah	Mastodon	0.036

To obtain a prior for θ, the pairwise divergence between within a population was calculated for all populations with unphased data. Sites with ambiguity codes were considered to be heterozygous in the individual and not due to sequencing error. As before, concatenated sequences were used and all gaps were removed prior to calculating the pairwise divergence. The μ prior was chosen to have a mean of 5×10^{-9} based on the priors and justifications used in previous analyses of this dataset (1).

Species	pairwise divergence
Asian	0.0012
Forest	0.0024
Mammoth	0.0008
Savannah	0.0006

B. Priors for mitochondrial dataset. The θ prior was determined by calculating the average pairwise divergence between all contemporary samples within a species across all possible pairs. Gaps were removed prior to calculating pairwise divergence. A relatively broad prior was chosen to reflect the large difference in average pairwise divergence in the different species.

Species	pairwise divergence
Asian	0.0034
Forest	0.013
Savannah	0.026

To find a prior for the root τ, the average pairwise divergence was found between all pairs of Asian and Forest elephants sequences and Asian and Savannah elephant sequences. The prior was chosen to have a mean close to the average pairwise divergence, with a relatively large variance to reflect the prior uncertainty in the parameter value.

Species 1	Species 2	pairwise divergence
Asian	Forest	0.047
Asian	Savannah	0.048

C. MCMC settings. MCMCs for the empirical analyses of both the nuclear and mitochondrial datasets were sampled 400,000 of times, sampling every 4 of iterations. The burnin was 160,000 iterations. Two and four independent MCMCs were run for each nuclear and mitochondrial dataset, respectively.

D. Convergence. Convergence was assessed in tracer by comparing the distributions of all parameters in the pairs of replicate MCMCs and examining the trace plot.

Accession No.	Species	Age
KY616982.1	*Loxodonta africana*	modern
KY616977.1	*Loxodonta africana*	modern
KY616974.1	*Loxodonta africana*	modern
AB443879.1	*Loxodonta africana*	modern
MT636097.1	*Loxodonta cyclotis*	1533 (417 ybp)
MT636095.1	*Loxodonta cyclotis*	1533 (417 ybp)
MT636093.1	*Loxodonta cyclotis*	1533 (417 ybp)
KY616981.1	*Loxodonta cyclotis*	modern
KY616980.1	*Loxodonta cyclotis*	modern
KY616975.1	*Loxodonta cyclotis*	modern
KJ557423.1	*Loxodonta cyclotis*	modern
NC_020759.1	*Loxodonta cyclotis*	modern
DQ316068.1	*Elephas maximus*	modern
OP575307.1	*Elephas maximus*	modern
OL628830.1	*Elephas maximus*	modern

Fig. S1. Additional samples used in the mitochondrial analysis downloaded from GenBank.

References

1. N Rohland, et al., Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biol.* **8**, e1000564 (2010).

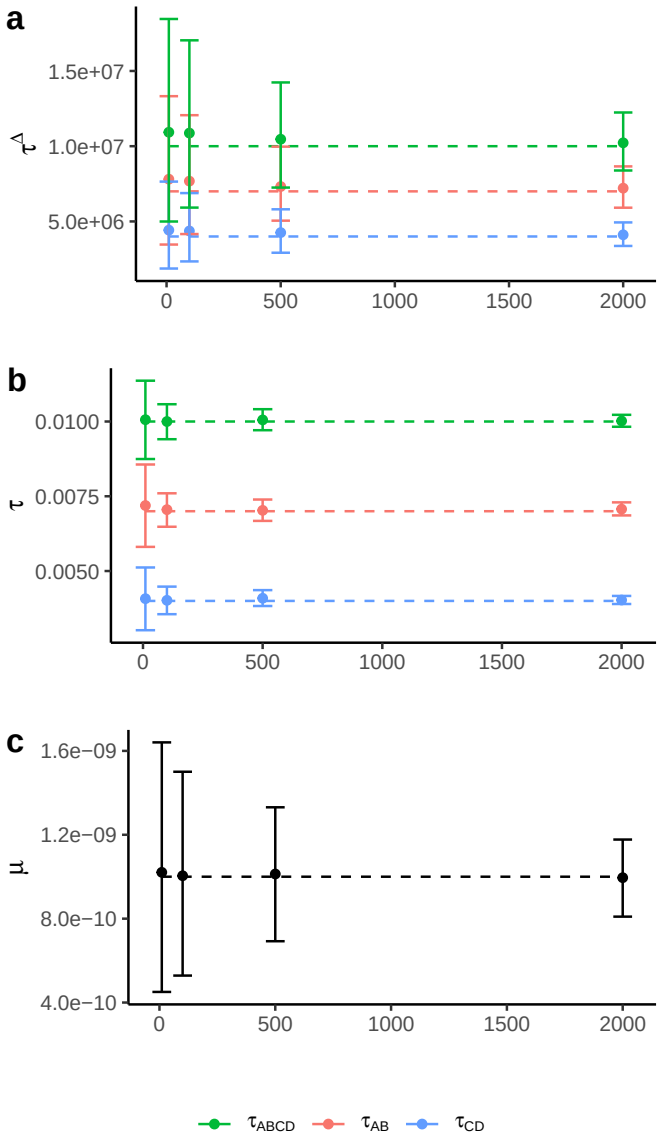

Fig. S2. Average posterior means and 95% HPD CIs (bars), over 20 replicate datasets, of (**a**) divergence times in mutations, (**b**) divergence times in years, and (**c**) mutation rate. The data were simulated under the model of figure 1a with two extinct species (A and C), sample dates are between 5,000 and 50,000 years, and $\theta = 0.001$.

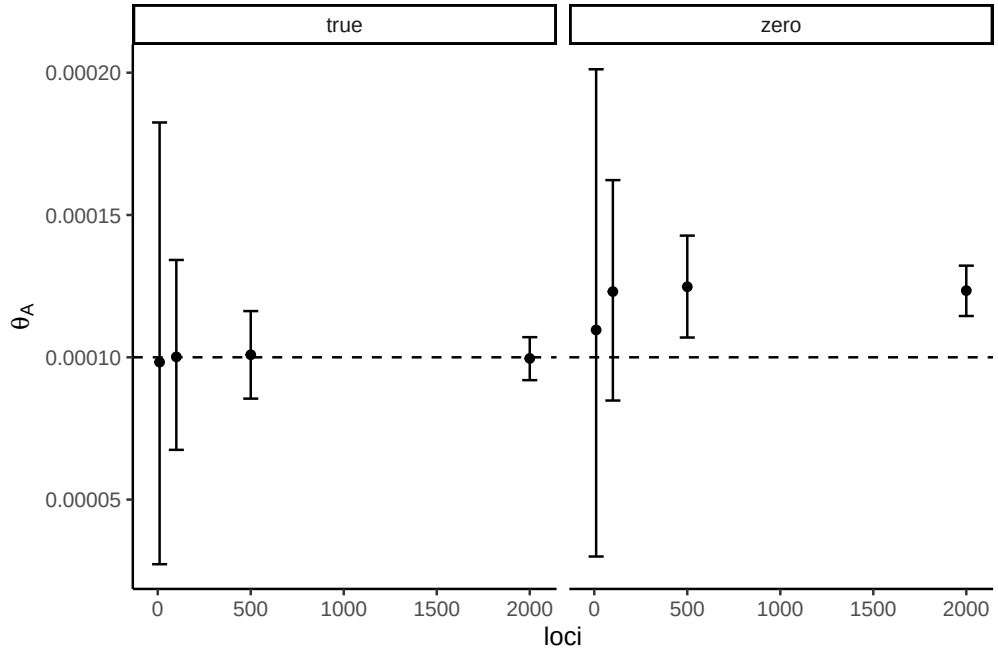

Fig. S3. Average posterior means and 95% HPD CIs (bars), over 40 replicate nuclear datasets, of θ_A when sample dates were set to their true values (left) or zero (right). The datasets had 6 samples in each extinct species and the upper bound on the sample dates equal to 50,000 ybp. The dashed lines show the true values of the θ_A.

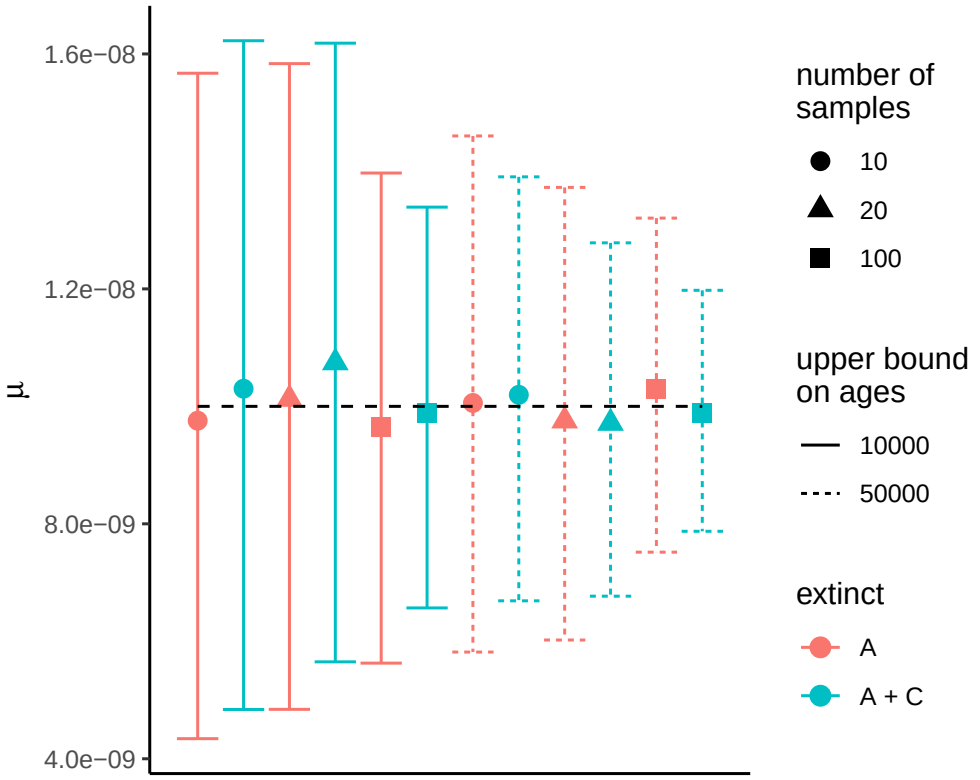

Fig. S4. Average posterior means and 95% HPD CIs (bars) of the mutation rate over 20 replicate datasets, simulated under the model of figure 1a with $\theta = 0.00025$. Solid lines are for sample dates between 5,000 to 10,000 ybp while dashed lines are for sample dates between 5,000 and 50,000 ybp. Either species A (red) or both A and C (teal) are extinct, and from each extinct species either 10 (circle), 20 (triangle), or 100 (square) samples are taken. The dashed line shows the true values of the μ.

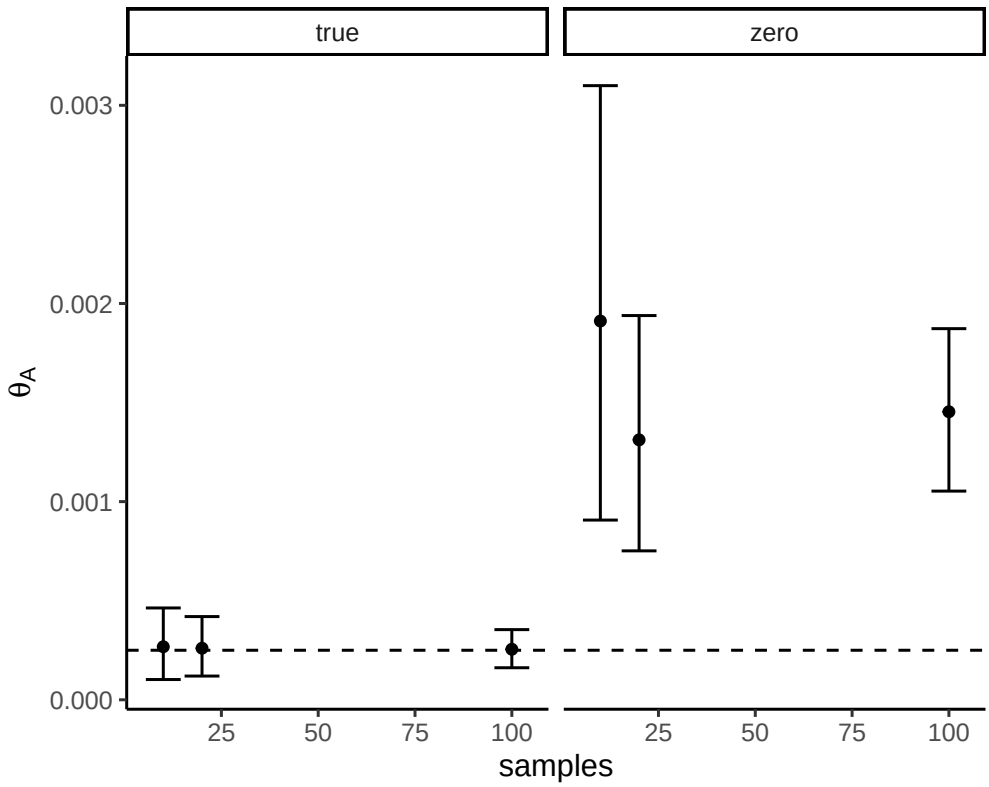

Fig. S5. Average posterior means and 95% HPD CIs (bars), over 40 replicate mitochondrial datasets, of θ_A when sample dates were set to their true values (left) or zero (right). The datasets had the upper bound on the sample dates equal to 50,000 ybp. The dashed lines show the true values of the θ_A.

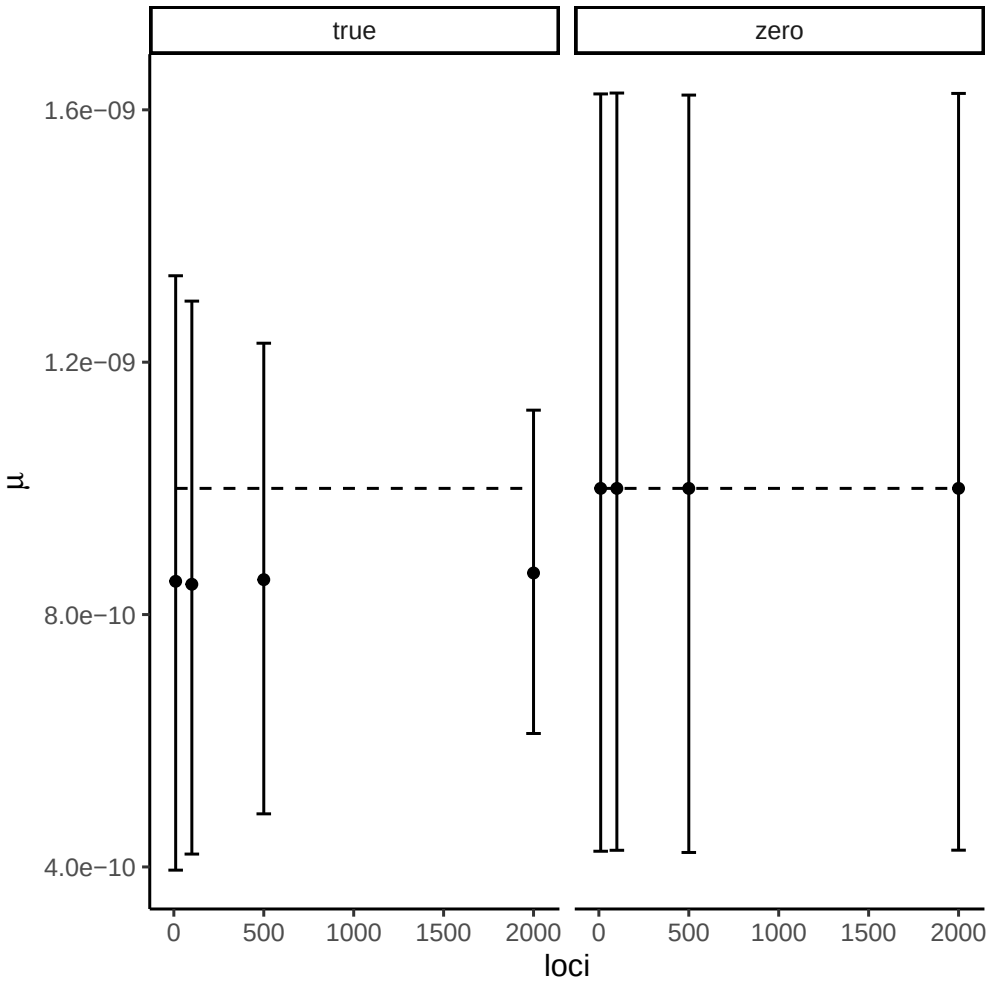

Fig. S6. Each point shows the average of the posterior mean mutation rate and mean 95% credible set averaged across inferences for 20 replicate datasets when the samples ages are set to their true value (left) or zero (right). For all datasets, there were 2000 loci and θ is equal to 0.0001. The dashed line shows the true value of μ.

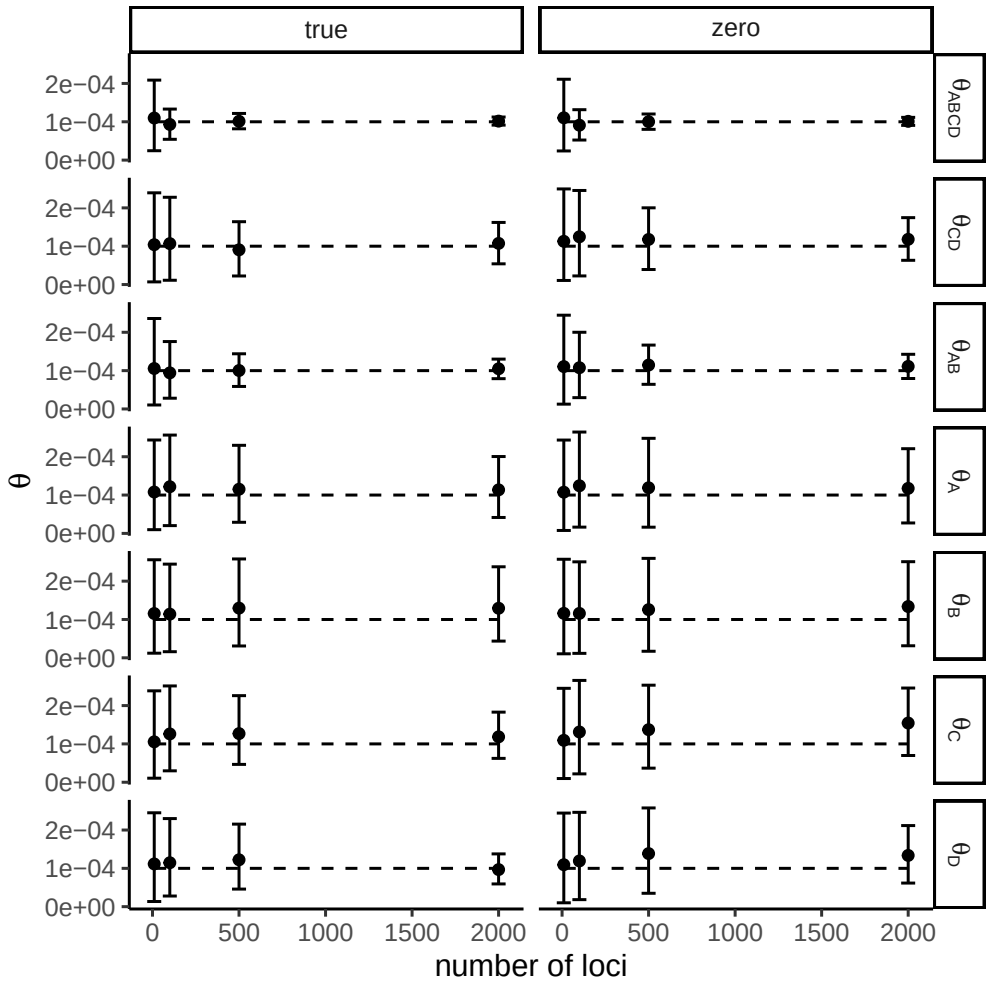

Fig. S7. Average posterior mean and 95% HPD CIs (bars) for θ across 20 replicate simulations for the recent population divergence analysis. The left and right plots show the inferences when the sample ages are set to their true values and zero, respectively. The dashed lines show the true value of θ.