**Title**

# Exploring Approximate Bayesian Computation for inferring recent demographic history with genomic markers in non-model species

**Authors**

Joane S. Elleouet, Sally N. Aitken

**Address**

Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia, 3041-2424 Main Mall, Vancouver BC V6T 1Z4, Canada

**Keywords**

approximate Bayesian computation, demographic inference, coalescent simulations, spatial expansion, population genetics

**Corresponding author**

Joane S. Elleouet, 3041-2424 Main Mall, Vancouver BC V6T 1Z4, Canada, joane.elleouet@alumni.ubc.ca

**Running title**

Inferring recent population history with ABC

## Abstract

Approximate Bayesian computation (ABC) is widely used to infer demographic history of populations and species using DNA markers. Genomic markers can now be developed for non-model species using reduced representation library (RRL) sequencing methods that select a fraction of the genome using targeted sequence capture or restriction enzymes (genotyping-by-sequencing, GBS). We explored the influence of marker number and length, knowledge of gametic phase, and tradeoffs between sample size and sequencing depth on the quality of demographic inferences performed with ABC. We focused on 2-population models of recent spatial expansion with varying numbers of unknown parameters. Performing ABC on simulated datasets with known parameter values, we found that the timing of a recent spatial expansion event could be precisely estimated in a 3-parameter model. Taking into account uncertainty in parameters such as initial population size and migration rate collectively decreased the precision of inferences dramatically. Phasing haplotypes did not improve results, regardless of sequence length. Numerous short sequences were as valuable as fewer, longer sequences, and performed best when a large sample size was sequenced at low individual depth, even when sequencing errors were added. ABC results were similar to results obtained with an alternative method based on the site frequency spectrum (SFS) when performed with unphased GBS-type markers. We conclude that unphased GBS-type datasets can be sufficient to precisely infer simple demographic models, and discuss possible improvements for the use of ABC with genomic data.


## Introduction

Patterns of DNA variation among individuals are commonly used to unravel events in the history of populations, such as demographic expansion, population splits, and admixture. Rapid progress in sequencing technologies at the start of the 21$^{st}$ century has allowed the inference of increasingly

complex demographic models, by using increasingly complete genomic datasets. However, this increase in amount of data and complexity of demographic scenarios necessitates new statistical methods for analysis and inference. Tackling large genetic datasets with inherent errors and uncertainties requires sophisticated techniques for marker development. In parallel, inferring complex historic demographic scenarios with several populations and numerous demographic parameters necessitates efficient algorithms to provide accurate parameter estimates and model validation measures. Reviews and improvements of methods have recently emerged (Schraiber & Akey, 2015), illustrating the fast pace of change in the field of statistical genetics. However, the efficiency of inference methods for different types of demographic models as well as effects of completeness of genomic datasets need to be understood to ensure quality and accuracy of inferences.

**Demographic inference in natural populations of non-model organisms**

In less than 30 years, human demographic inference has taken a leap, evolving from the evidence for a single African origin of all humans using a few non-recombining mitochondrial markers (Cann et al., 1987), to the inference of highly complex demographic scenarios using whole genomes (Harris & Nielsen, 2013). Although there is still room for improvement in demographic inference of human populations (Schraiber & Akey, 2015), human genomics is at the leading edge of inference from DNA data. Unfortunately, the state-of-the-art statistical inference techniques applied to human data are currently out of reach for studies of natural populations of non-model organisms. Knowledge from demographic inference of these species is, however, crucial: it is often the most efficient way to determine how to manage invasive species (Benazzo et al., 2015; Guillemaud et al., 2010), to conserve endangered species or ecosystems (Chan et al., 2014; Dussex et al., 2014; Lopez et al., 2006; Quéméré et al., 2012), and to predict the future distribution and abundance of widespread species that are of

economical or ecological importance (Holliday et al., 2010; Zinck & Rajora, 2016). The good news is the genomic revolution has reached non-model organisms, creating a spectrum of levels of genetic knowledge across a broad range of taxa. Using a few microsatellites or moderate-sized panels of resequenced SNPs is still common practice (Y. Li et al., 2010; Zinck & Rajora, 2016), but most current studies of non-model species now use genomic methods to extract markers for inference. In recent years, sequencing whole genomes of non-model species has become feasible in some organisms with small genomes (Boitard et al., 2016; Liu et al., 2014) and has allowed the inference of detailed demographic models using Approximate Bayesian Computation (ABC) or Pairwise Sequential Markovian Coalescent (PSMC) (Nadachowska-Brzyska et al., 2013). For organisms with larger genomes or for studies with lower data requirements, reduced-representation library (RRL) sequencing, through either targeted capture or restriction enzymes, is widely applied (Davey et al., 2011). RRL techniques involving restriction enzymes (commonly referred to as RADseq or genotyping-by-sequencing, GBS) output a large number of short sequences (100bp, or longer with paired-end sequencing) from across the genome and have proven useful in population genetics studies and inference involving maximum likelihood methods based on the site frequency spectrum (SFS) or ABC methods (Narum et al., 2013). Most recently, the number of published drafts of whole genomes for non-model species has increased dramatically, granting access to longer sequences through the second category of genomic markers: targeted enrichment. This approach allows the use of linkage information for population genetics inference (Li & Jakobsson, 2012).

**Approximate Bayesian Computation and other approaches**

In this paper, our aim is to explore ABC for datasets obtained from reduced-representation library sequencing in non-model organisms. We also compare the results obtained with those from a SFS

approach based on approximation of the composite likelihood (Excoffier & Foll, 2011). We chose to explore ABC because of its versatility: It accommodates a wide spectrum of demographic models and dataset types. Although it was originally developed for inferences in evolutionary biology, the statistical framework of ABC has been extended to a variety of disciplines, from cell biochemistry and epidemiology to neural networks, extending beyond the realm of biology into meteorology, astrophysics (Weyant et al., 2013) and computer sciences (Condon & Cukier, 2016). ABC has been reviewed in a number of publications and its algorithms and techniques are being refined constantly (Bertorelle et al., 2010; Csilléry et al., 2010; Lintusaari et al., 2016; Marin et al., 2012; Sunnaker et al., 2013). For applications in demographic inference using genetic data, the general ABC method involves the following steps. First, a large number of datasets are simulated under a specific demographic model using the coalescent (Kingman, 1982). Parameters used for simulations are drawn from prior distributions that are pre-defined by the user. The simulated datasets are then compared to the observed dataset through calculation of summary statistics. Finally, simulated datasets with the closest vector of statistics to the vector of observed summary statistics are selected. A regression adjustment based on the local relationship between statistics and parameters is then usually performed to approximate the posterior distribution of each model parameter from the parameter values of selected simulations. ABC is suitable when inferring models for which the likelihood function is intractable, as it relies on approximating the likelihood function using a large number of simulations. However, each one of the numerous steps in the implementation of ABC requires users to make empirical decisions. There is particularly a need to improve our understanding of the relationship between the type of markers obtained to build genetic datasets and the way genetic data is subsequently summarized on its power to tease apart demographic models and produce accurate parameter estimates.

**Previous work exploring ABC**

The need to test the inference power of datasets for demographic models of interest has been recognized in recent years, both in terms of model selection and parameter estimation. Robert et al. (2011) warned against the use of insufficient summary statistics in ABC model choice, opening the door to improved methods for model testing and the associated choice of summary statistics (Marin et al., 2014; Prangle et al., 2013). Among theoretical results and general guidelines, Marin et al. (2014) suggested the use of different sets of summary statistics for estimation and model selection. Several studies show the use of preliminary simulations testing parameter estimation and model choice with different number and length of markers and number of individuals (Sousa et al., 2012; Stocks et al., 2014), type of molecular markers (Cabrera & Palsbøll, 2017) and choice of summary statistics and models considered (Benazzo et al., 2015; Guillemaud et al., 2010; Li & Jakobsson, 2012; Sousa et al., 2012; Stocks et al., 2014). As most scientists have switched to using genome-wide data, there is a need to expand this set of simulation studies to test and understand the power of different types of genomic data. As part of such an effort, Li & Jakobsson (2012) simulated large, phased genomic datasets comparable to human genomic datasets at the time. Under 2-population split models, they found that ABC produces accurate estimates for most but not all parameters and concluded ABC is well suited to large genomic datasets summarized with LD-based statistics. Robinson et al. (2014) tested the effects of the number and length of unphased genomic sequences and compared them to the effect of the number of individuals sequenced for the inference of three-population admixture models. They found that increasing the number and length of sequences was more beneficial than increasing sample size. Shafer et al. (2015) investigated the power of ABC on short diploid sequences obtained by GBS. They focused on a wide range of simple 1-population and 2-population models with bottleneck, growth, migration and a combination of these parameters. They found that population changes such as ancient

136 temporary bottlenecks would not be inferred correctly regardless of the number of markers available.
137 This set of studies provides valuable information about the use of genomic data in ABC. Our aim is to
138 extend this knowledge by directly comparing ABC results from molecular markers obtained with
139 different types of RRL sequencing techniques, different sequencing effort allocations, and different
140 levels of genomic knowledge. This will hopefully help future ABC users who do not have access to
141 complete genomic data to select methods and develop genomic datasets that are best suited to answer
142 the demographic questions they are addressing.
143
144 **General model and datasets**
145 Here, we focused on estimating parameters for a set of 2-population models of demic expansion
146 that are applicable to studies of species invasion, reintroduction, or natural colonization. We tested the
147 power of ABC on these models using a range of marker sets obtainable by RRL methods: datasets with a
148 large number of short genomic reads would correspond to single-end GBS sequencing, whereas fewer
149 but longer diploid sequences correspond to a targeted enrichment approach. For each type of dataset,
150 we quantified the potential benefits of knowing the gametic phase of sequence markers by including or
151 excluding linkage-related statistics at the data-summarizing step. We expect to observe an improvement
152 in the inference for datasets with long sequences. For each model assessed, we also tested the effect of
153 time since colonization. We hypothesize that recent events might be inferred more accurately with
154 datasets containing linkage information, due to the generally higher rate of recombination compared to
155 mutation, and to the potential information contained in long haplotypes. This part of the analysis is also
156 motivated by the fact that overestimates of divergence times are a common result of demographic
157 inference in empirical studies (Holliday et al., 2010) and this upward bias has been found for some
158 demographic scenarios in simulation studies (Benazzo et al., 2015). We therefore aim to explore this

159 potential bias by testing increasingly old events within the same models. As NGS techniques require a
160 trade-off between sample size and individual sequencing depth, and are characterized by high
161 genotyping errors, we explore the effect of different trade-offs at different sequencing error rates.
162 Fumagalli (2013) found that increasing sample size at the cost of decreasing depth was beneficial in the
163 inference of diversity measures and population structure. Here, we extend this hypothesis to ABC
164 inference. Finally, we compared our ABC results with those obtained from an approximate likelihood
165 method using the site frequency spectrum from simulated reduced-representation libraries. As they
166 provide millions of genome-wide SNPs without ascertainment bias, restriction enzyme-based genomic
167 sequencing techniques seem to be particularly well suited to SFS-based inference methods. Comparing
168 SFS results with ABC results on a range of models and datasets will inform future work on demographic
169 inference in non-model organisms.
170

171 # Methods

172 ## Demographic models

173 We focused on a basic 2-population model of demic expansion (fig.1a). A pre-existing
174 population, population 1, is of constant size $N_1$. At time $T_{EXP}$ before present, the spatial population
175 expansion begins: population 2 is created by 2 migrants from population 1. Population 2 then grows
176 exponentially between times $t=T_{EXP}$ and t=0 (the present) to size $N_2$ at t=0. The rate of population growth
177 $r$ is defined by the other parameter values through the formula $r=log\ (\frac{N_{02}}{N_2})/T_{EXP}$. Model 1 therefore has
178 just 3 independent unknown parameters: $N_1$, $N_2$, and $T_{EXP}$. We created additional models of increasing
179 complexity by adding parameters. In models 2 and 4, the number of founders of population 2, $N_{02}$, is
180 unknown (fig.1b and fig.1d); in models 3 and 4, migration is allowed from population 1 to population 2,
181 with the parameter $m_{21}$ describing a per-generation migration rate (fig.1c and fig.1d). In all four models

182 described above, the mutation rate and the recombination rate are fixed. We chose wide and uniform
183 parameter priors for population sizes to accommodate a wide range of types of organisms, and a log-
184 uniform prior for the timing of the expansion event, as this study intends to focus on more recent rather
185 than ancient expansion events (Table 1).
186
187 **Generating sets of coalescent simulations**
188 For each of the four models, we created a set of 1 million simulations with each of the five types of
189 datasets described below, with a fixed number of 10 diploid individuals sampled per population. For
190 datasets corresponding to single-end RADseq sequencing techniques, we simulated 10,000 independent
191 DNA sequences of 100bp each. For datasets corresponding to sequence capture methods, we created
192 100 independent DNA sequences of 10kb each. Additionally, we explored a range of possible
193 configurations between these two types of datasets (Table 2). With 4 models and 5 types of datasets, we
194 obtained a total of 20 combinations of models and datasets, each with a million simulations. We used
195 the program scrm (Staab et al., 2015), which simulates datasets by creating the ancestral recombination
196 graph following the Wiuf and Hein method (1999). We used custom Rscripts (R Core Team, 2016)
197 inspired by scripts from Shafer et al. (2015) to compute the simulations, and made them available in the
198 supporting information.
199
200 **Summary statistics**
201 For each simulation we computed all summary statistics available in the program msABC
202 (Pavlidis et al., 2010). The available statistics include diversity statistics (number of segregating sites and
203 $\theta$ estimates) and summaries of the SFS (Tajima's D and Fay and Wu's H). These statistics were calculated

for each population and for the whole sample. The available statistics also include summaries of the 2d-SFS: differentiation measures such as the pairwise $F_{ST}$ and the number of private and shared polymorphisms. Finally the Thomson estimator of $T_{MRCA}$ and its variance were calculated for each population and for the whole sample. To test the effect that knowing haplotype information has on inference, the ABC analysis was performed twice on each model-dataset type combination. The first time, we summarized data using only the statistics mentioned above, which are calculated at the SNP level and therefore are available when the gametic phase of the diploid sequences is unknown. The second inference was performed on the same dataset, but additional statistics Zns (Kelly, 1997), dvk and dvh (Depaulis & Veuille, 1998) based on linkage information were used to summarize the data. These additional statistics are calculated at the haplotype level and so are only available in cases where the gametic phase of the diploid sequences is known. For each set of simulations, we computed the mean and variance of every statistic over all sequence markers in the dataset. As a result, 58 statistics were computed for datasets with known gametic phases (hereafter referred to as "phased", or "hap.phase 1"), and 43 statistics were computed for datasets with unknown gametic phases (hereafter referred to as "unphased", or "hap.phase 0").

Using a high number of statistics to summarize genetic data has harmful effects on the quality of the ABC inference, a problem commonly referred to as the "curse of dimensionality" (Blum et al., 2013). We used the partial least squares (PLS) method implemented in ABCtoolbox (Wegmann et al., 2010) to reduce the number of statistics to 5-7 PLS components (see Supplemental methods for details).

**Pseudo-observed datasets**

For each set of 1M simulations, we created a corresponding set of 100 pseudo observed datasets (PODs), with parameters randomly chosen from the same priors as for the set of 1M simulations. By doing so we assume that priors are reliable and reflect the true, unknown distribution of the PODs. These were then summarized with the same summary statistics as their corresponding set of 1M simulations.

**ABC estimation**

We performed the ABC estimation using each POD as the observed dataset to obtain parameter estimates. The standard ESTIMATE algorithm from the program ABCtoolbox (Wegmann et al., 2010) was used for all ABC computations to create posterior probabilities from the corresponding set of 1M simulations, with a post-sampling regression adjustment through ABC-GLM (Leuenberger & Wegmann, 2010). We fixed the tolerance parameter to $10^{-3}$, a compromise between having a tolerance threshold value as low as possible (Li & Jakobsson, 2012) and keeping an appropriate number of simulations to estimate the posterior from.

**Validation**

For each combination of model and type of dataset, we computed a measure of precision and accuracy called the relative prediction error (RPE), the ratio of the mean squared error over the variance of the prior, which follows equation (2):

$$(2)\ \mathcal{E} = \frac{\sum_{j=1}^{j=i}(\widehat{\theta_J} - {\theta_j}^*)^2}{Var(\theta)} \times \frac{1}{i}$$

where $Var(\theta)$ is the variance of the prior distribution and i is the number of observations. The RPE was computed on 1,000 PODs. The advantage of using RPE as a validation statistic is that it directly indicates the contribution of the genetic dataset to the estimation of the posterior. Another attractive feature of the RPE is that it allows comparisons between parameters, as it scales from 0 (precise estimate) to 1 and beyond (in the case of a consistent bias in estimation).

As an additional measure of precision, the 95% highest posterior density interval (HDI) was calculated on a set of 100 PODs for each combination of model and dataset type. This measure is defined as the shortest continuous interval with an integrated posterior density of a certain value (Wegmann et al., 2010). For each combination of model and dataset type we reported the 95% HDI coverage, i.e. the number of times (out of 100) the true parameter value fell within the 95% HDI, expecting values close to 95.

**Testing the effect of $T_{EXP}$ on parameter estimation**

To test the effect of the time of expansion on the precision of the ABC estimation, we created 100 PODs for each set of 1M simulations and 12 fixed values of $\log T_{EXP}$ spanning the prior range. RPE and 95%HDI were calculated from the results of each set of 100 PODs.

**Effect of sequencing effort allocation and sequencing error**

The main challenge when developing genomic markers is managing sequencing and variant calling errors. Sequencing a large number of individuals might increase the precision of population genetics inference, but with a fixed sequencing budget, this comes at the cost of reduced individual

sequencing depth, which in turn can affect variant calling and estimation of allelic frequencies (Fumagalli, 2013). We explored this challenge focusing on model 2 and dataset type 2. We chose a realistic fixed sequencing effort and derived 3 fixed sampling strategies from it: 250 sampled individuals at a mean individual depth of 4, 100 individuals with depth 10, and 20 individuals with depth 50. We then incorporated three per-nucleotide sequencing error rates (0, $10^{-2}$, $10^{-3}$), and applied them to each category described above. The resulting 9 categories of PODs, as well as "perfect" datasets (no depth sampling and no error) were all simulated using the same 10 parameter combinations. Further details about the creation of "imperfect" PODs can be found in the supplemental methods. Once these imperfect PODs were created and summarized, ABC was performed to estimate their true parameter values. Two additional sets of 1M simulations needed to be created to match the number of individuals sampled per population: one with 100 diploids per population, and the second with 250. It the latter case, we only created 610,000 simulations because of computation time limitations. The same tolerance (0.001) as all other runs was used for the estimation.

### Comparing ABC and SFS estimation

We simulated 10,000 independent DNA sequences of 100bp each for the 4 demographic models 10 times. The resulting 40 datasets were input into both ABCtoolbox and fastsimcoal2, which uses the SFS to approximate a composite likelihood from a large number of simulations through a conditional maximization algorithm (see supplemental methods). We compared the results from the two methods using RPE, credible intervals and confidence intervals.

## Results

A total of 20 combinations of models and datasets were used as input for ABC simulations (Tables 1 and 2), resulting in a total of 20 million simulated datasets available for analysis, training simulation sets and PODs. Each set of 1M simulations was used in two runs of estimation: one including all summary statistics available in msABC, the other one excluding statistics based on linkage information, for a total of 40 ABC estimations.

**Effect of model complexity on the precision of parameter estimates**

In general, the ability to infer demographic history declined rapidly as model complexity increased. The simplest model (1), estimating only population sizes $N_1$ and $N_2$ and the log-transformed time of expansion $T_{EXP}$, allowed the expansion event to be dated accurately. Models 2 and 3 each had 4 parameters: model 2 included the number of founders $N_{02}$ and model 3 allowed migration from population 1 to population 2 ($m_{21}$). For both model 2 and 3, $\log T_{EXP}$ was inferred with slightly lower precision than for model 1. Finally, scenarios corresponding to model 4, which had all 5 parameters, failed to be correctly inferred.

Not all parameter estimates were sensitive to the addition of parameters in the models: the precision of contemporary population size estimates $N_1$ and $N_2$ were independent of model complexity. RPE values for $N_1$, which was constant over generations, were mostly below 0.05 for the four models assessed (fig. 2). The 95% highest posterior density intervals ranged from 3,000 to 60,000. For $N_2$, the contemporary population 2 size after exponential growth, 95% HDI intervals were about as wide as the prior range, indicating a failure to estimate this parameter in all four models (fig. 3).

The expansion time $T_{EXP}$ was generally well estimated in model 1, which is the simplest 3-parameter model (fig. 2) with no migration between demes and the number of founders set to 2. For

this model, the RPE was mostly below 0.1. The precision of $\log T_{EXP}$ estimation was almost as high for the two 4-parameter models, where the number of founders $N_{02}$ (model 2) is unknown and needs to be estimated, or where migration from population 1 to population 2 is likely (model 3). For these two models, the RPE is below 0.2. The ABC analysis of the 5-parameter model (model 4) was unable to recover the true $T_{EXP}$ value.

Estimates of the number of founders of population 2 ($N_{02}$) and migration rate from population 1 to 2 ($m_{21}$) were surprisingly imprecise in models of low complexity (model 2 and 3) and could not be recovered at all in model 4 (fig.2 and 3).

Models 1 to 4 all rely on population 2 growing exponentially from $T_{EXP}$ to the present time. We tested whether demographic parameters could be estimated more successfully in a model where population 2 goes through a single sudden population change instead of exponential growth. We created a new set of 1M simulations based on model 2 (where $N_{02}$ is a varying parameter) and dataset type 1 (many short sequences) and a smaller prior range for $T_{EXP}$ (2-500 generations). In the new model the size of population 2 changes from $N_{02}$ to $N_2$ at $T_{EXP}/10$ and remains constant before and after $T_{EXP}/10$. These modifications brought no improvements to any of the parameter estimates (Table S1).

**Do sequence length and linkage-related statistics improve the estimation?**

The addition of linkage statistics available in msABC brought no notable improvement in the RPE and 95% HDI of parameter estimates for all models (fig.2 and fig.3). It even seems to make the estimation of $N_1$ less precise in some cases for model 1, 2 and 4, although this pattern is inconsistent across dataset types. ABC performance on models 3 and 4 seemed to be slightly more dependent on

sequence length, with the inference on large sequences marginally benefitting from haplotype information.

**Quality of parameter estimates across prior ranges**

For each parameter, we visualized estimated values and 95% HDI of ABC results in relation to true parameter values to assess performance over the prior range. Results for the 3-parameter model (model 1) and dataset types 1 and 5 are shown in fig. 4a and 4b, respectively. Results for the complete set of models are available in supplemental fig. S1. Consistently across models, estimates of $N_2$, $N_{02}$, and $m_{21}$ are largely inaccurate regardless of the true value, with HDI ranges as wide as the prior range. Conversely, $N_1$ estimates are accurate in all models regardless of the true $N_1$ value. Unlike $N_1$, the values of $T_{EXP}$ have an impact on the precision of their respective estimates. Accuracy and precision of $T_{EXP}$ estimates for models 1 and 3 decrease with increasing true value. Interestingly, the opposite pattern is observed for model 2: more recent events are less precisely inferred than ancient ones (fig. S1, pp. 11-20). Results for model 4 show a "cross" pattern where most PODs' $\log T_{EXP}$ values are correctly estimated but some PODs with extreme $\log T_{EXP}$ values show estimates at the opposite extreme (fig. S1, pp. 31-32). This pattern suggests a complex multivariate relationship between model parameters and statistics.

**Effect of the time of the expansion event on the estimation**

We tested whether older expansion events are generally more difficult to characterize than recent ones within the time range specified by the prior. To do this, we studied the effect of the true $T_{EXP}$ value on the precision of parameter estimates. We find different trends among the 4 models (fig. 5, S2,

and S3). The precision of inference on model 1 is higher at low $T_{EXP}$ values and decreases at $\log T_{EXP}>4$. Conversely, for model 2, older events are generally better inferred: estimates of $T_{EXP}$ and $N_{02}$ increase in precision as $T_{EXP}$ increases, as shown by the RPE (fig. S2, p.2) and the 95% HDI (fig. S3, p.2). Model 3 shows the best results for moderately recent expansion events ($3 < \log T_{EXP} < 4$), as shown by RPE and 95% HDI of $T_{EXP}$ and $m_{21}$ (fig.S2 and S3). Finally, results for model 4 show high values of RPE and 95% HDI for all parameters, with RPE values mostly above 0.5.

**Effect of sequencing effort allocation and sequencing error**

Focusing on model 2 and datasets of 5,000 x 200bp sequences, we simulated sequencing and variant calling for three different sample size and depth combinations. The RPE of parameter estimates for 13 tested PODs is represented in fig. 6. Depth of sequencing (dp) has very little effect on the precision of estimates: only $N_1$ and $\log T_{EXP}$ have a marginally higher RPE when sequencing depth is simulated. Error rates affect $N_{02}$ estimates at low depth (N=250, dp=250), as well as $\log T_{EXP}$ estimates at low sample size (N=20, dp=50). The estimation is otherwise robust to introduced errors. For a given set of PODs (e.g. N=250, dp=4), the precision lost in a parameter estimate because of an error rate of 0.01 ($N_{02}$) is gained on another parameter ($N_1$), reflecting the limitations of the model estimation process rather than the effect of sequencing error. However, the results suggest that choosing a larger sample size with a shallower individual sequencing depth improves estimation over other strategies, especially for the estimation of $\log T_{EXP}$.

**Comparing ABC with SFS estimation using an approximate composite likelihood**

373 Figures 7 and 8 illustrate the performance of ABC and approximate composite likelihood from
374 the SFS for all models performed with datasets of 10,000 100-kb sequences. Both methods gave similar
375 results in terms of precision of parameter estimates. The SFS-based method performed slightly better
376 than ABC in the model with migration (model 3), but the precision of ABC estimates was superior for
377 model 2 (fig.7). The approximate composite likelihood method generally provided narrower 95%
378 confidence intervals (fig.8).

379

## 380 Discussion

381 We explored the ability of approximate Bayesian computation to characterize a recent event of
382 spatial expansion from one population of constant size to a new and growing population, a model which
383 can be broadly applied to studies of species range expansion, invasion biology, or reintroduction of
384 endangered species. We found that regardless of model complexity, estimates of the size of the growing,
385 newly founded population ($N_2$) are poor. However this did not prevent successful estimation of other
386 parameters ($N_1$, $\log T_{EXP}$, and in restricted cases $N_{02}$). Failure to estimate $N_2$ does not come as a surprise:
387 estimates of past changes in effective population size from one punctual sampling event commonly rely
388 on linkage information between markers, a calculation not readily available in ABC packages (Beaumont,
389 2003). Our result that models of higher complexity are harder to estimate was expected, but in the case
390 of our expansion models, this trend leads surprisingly quickly to a complete failure to estimate any
391 parameter, as soon as 5 parameters are involved. While expansion timing was precisely estimated in the
392 3- and 4- parameter models, it could not be recovered in the 5-parameter model. ABC on model 2, the 4-
393 parameter model including the number of founders but no subsequent migration, successfully estimated
394 all parameters (except $N_2$) for old expansion events. In contrast, for model 3, the 4-parameter model

395 including migration between demes, estimations were more successful for recent events. These results
396 highlight the potential importance of taking into account the timing of an expansion event when
397 predicting estimation success for a given demographic model. The difficulty of estimating the time of a
398 founding event with subsequent migration was also reported by Robinson et al. (2014); however, we
399 show here that for a moderately recent event (10 to 100 generations), it is possible.
400
401 **Implications of including haplotype information**

402 Analyses based on unphased sequences exploring similar models to those used here have shown
403 encouraging results (Robinson et al., 2014). However, no study to date has explicitly compared datasets
404 of phased and unphased sequences using the same models and same amount of data. Here, we
405 quantified the benefits of using phased haplotype sequences over single SNPs by including or leaving out
406 LD-based and haplotype-level statistics at the data summarization step of the ABC inference.
407 Surprisingly, haplotype information did not substantially improve the precision of parameter estimates,
408 even when 10-kb sequences were used as markers. Li and Jakobsson (2012) explored ABC with similar 2-
409 population split models and a similar fixed population-wise per-generation recombination rate as in our
410 study. When they tested different combinations of summary statistics, their results did not demonstrate
411 any obvious superiority of LD-based statistics over SNP-based statistics. They concluded that the selected
412 summary statistics should capture as many different aspects of the data as possible, with as little
413 redundancy as possible. Potentially, phasing the data may not have improved inferences because the
414 extent of linkage that the chosen statistics are sensitive to differs from the linkage actually present in the
415 simulated data. Future work when dealing with phased data would require developing expectations of
416 LD levels and creating or choosing statistics that cover the extent of LD likely to be present in the data.

One needs to be aware of the difficulties associated with the use of LD information. Firstly, ABC on phased data requires reasonable knowledge of recombination rates and variability across the genome. The recombination rate needs to be included as a parameter along with demographic parameters, or as a nuisance parameter with a hyper-prior. Secondly, simulating the coalescent with recombination is a complicated process and comes at high computational costs (McVean & Cardin, 2005). With high recombination rates or very long sequences, coalescent simulations might take so long to run that one would instead use a more efficient inference method than ABC. Moreover, translating genome-wide observed data into a set of summary statistic values that are readily useable by ABC programs and comparable to simulated datasets can be a challenge. File input formats in most programs are currently not compatible with sequence information, and many summary statistics programs do not offer haplotype-level calculations. Thirdly, when aligning reads to a fragmented and incomplete reference genome, as is often the case for non-model organisms, defining haplotypes can be tricky. One also needs to address problems of sequencing errors, paralogous sequences and imperfect mapping. Inevitable sequencing uncertainties will affect haplotype statistics more strongly than single-SNP diversity measures. Data processing errors and filters can severely bias inferences, to the extent of supporting the wrong demographic model, as revealed by Shafer et al. (2016). Finally, targeted sequence capture will result in thousands of markers of various lengths. Setting up simulations that correspond closely to an observed dataset requires approximating the distribution of sequence lengths, and this may also affect inferences, especially if variances of summary statistics are included at the data summarization step. Considering the difficulty of obtaining reliable haplotype information in non-model organisms, the potential difficulties of adapting the use of long sequences to currently available ABC programs, and computational time, our results tend to suggest that using SNP-level information from GBS-type data is preferable over targeted sequence capture.

440

441 **Choosing summary statistics**

442 It is important to note that all the results presented here are only valid in the context of our
443 choice of summary statistics. In the present study, we decided to use the first and second moment of all
444 statistics available in msABC, and to reduce the dimensionality with a PLS transformation. Several
445 previous publications have performed simulations either using the two first moments of summary
446 statistics (Li & Jakobsson, 2012) or only using the mean (Shafer et al., 2015). To our knowledge, only
447 Robinson et al. (2014) tested the use of 4 moments for summary statistics for models of divergence with
448 admixture. They compared their results with those obtained using only the mean and found that the
449 mean alone was sufficient. Although the two first moments may not be the most representative
450 summaries for some statistics, adding higher-level moments will come at a computational cost.

451

452 It is widely recognized that choosing a set of summary statistics is probably the most challenging
453 step for ABC users. For instance, the optimal set of statistics for parameter estimation in a given model
454 might differ from the optimal set of statistics to discriminate between demographic models. As
455 insufficient summary statistics have detrimental effects on model selection (Robert et al., 2011),
456 Fernhead and Prangle (2012) introduced "semi-automatic ABC", which relies on an ABC pilot run and a
457 subsequent linear regression to choose the most appropriate set of summary statistics. Similarly,
458 ABCtoolbox 2.0 implements a statistical selection step based on the incremental assessment of inference
459 power with the addition of summary statistics. However, documentation is lacking for this new feature
460 of the program. These improvements constitute a promising step towards a more rigourous statistical
461 framework for the automatic selection of ABC summary statistics.

462

463 **Sequencing effort: go large and shallow!**

464 We found that "imperfect" datasets created with a high number of individuals sequenced at a
465 low individual depth seemed to perform consistently better for most parameters than datasets with
466 fewer individuals and higher depth. This is consistent with Fumagalli (2013), who studied the same
467 trade-offs on diversity statistics under various demographic settings. This result seems to hold even with
468 simulations with moderate or high sequencing error rates, although this is difficult to conclude with
469 confidence considering the large bootstrapped confidence intervals (fig.6). It is worth noting that if the
470 error rate is not properly estimated during the genotype calling process, more errors will be present in
471 the final dataset and it is likely that ABC results will be impacted for all sequencing strategies, especially
472 those with low depth. As ABC summary statistics rely on the SFS and not on individual genotypes, we
473 suggest that future ABC users sequence large sample sizes at low depth. In this case, estimating the SFS
474 or derived statistics following methods such as described in Nielsen et al. (2012) and Fumagalli et al.
475 (2014) has proven more successful than genotype calling in inferring the SFS. There is unfortunately no
476 straightforward program or pipeline of compatible programs incorporating these methods into an ABC
477 framework. One possibility is to summarize the SFS into quantiles and to use the latter as summary
478 statistics in a classic ABC run. Such a process would need to be further tested.
479

480 **Comparing ABC to other methods**

481 We did not find large differences in the precision of parameter estimates between ABC and the
482 SFS-based likelihood method implemented in fastsimcoal2. Shafer et al. (2015) found a similar result
483 while comparing the performance of ABC with a SFS-based inference implemented in *δaδi* (Gutenkunst
484 et al., 2009). They found that *δaδi* tends to overestimate the time of population split and bottleneck
485 events, a trend not supported by our findings with *fastsimcoal*. In addition to parameter estimation,

Shafer *et al.* (2015) tested the performance of both methods for model selection and found ABC more accurate, especially in the case of bottleneck scenarios.

ABC has proven moderately useful for demographic inference with long, genome-wide haplotypes but comparisons with alternative approaches are scarce. Notable examples include Nadachowska-Brzyska et al. (2013), who used ABC and PSMC in a complementary way. Robinson et al. (2014) compared their ABC results with an exact likelihood method developed by Lohse et al. (2011) and found that ABC resulted in more uncertainty, especially in model comparisons. As ABC performance with linkage information needs to be further explored, comparisons to emerging analytical methods based on whole genomes or long sequences such as MSMC (Schiffels & Durbin, 2014) or identity-by-descent haplotype sharing (Harris & Nielsen, 2013) will greatly help refine methods for demographic inference using data at a genomic scale.

Theoretical improvements of ABC methods are emerging rapidly. Although the results presented here do not show that ABC benefits greatly from the use of deeper genomic datasets, the versatility of ABC might be key to its useful applications in a wide variety of fields, even those progressing rapidly such as population genetics. Constant methodological improvement, however, requires regular updates to available ABC programs.

## Acknowledgements

J.E was supported by an NSERC Discovery Grant to S.N.A and a Strategic Recruitment Fellowship from the Faculty of Forestry, University of British Columbia. We thank Michael Whitlock for his insightful comments on the manuscript, and Daniel Wegmann for providing bioinformatics support. Thanks to the two anonymous reviewers for their helpful comments.

## References

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, *164*, 1139–1160.

Benazzo, A., Ghirotto, S., Vilaça, S. T., & Hoban, S. (2015). Using ABC and microsatellite data to detect multiple introductions of invasive species from a single source. *Heredity*, *115*, 262–272. https://doi.org/10.1038/hdy.2015.38

Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. *Molecular Ecology*, *19*, 2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x

Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*, 189–208. https://doi.org/10.1214/12-STS406

Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from large samples of genome-wide molecular data: an approximate Bayesian computation approach. *PLOS Genetics*, *12*, e1005877–e1005877. https://doi.org/10.1371/journal.pgen.1005877

Cabrera, A. A., & Palsbøll, P. J. (2017). Inferring past demographic changes from contemporary genetic data: A simulation-based evaluation of the ABC methods implemented in DIYABC. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.12696

Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*, 31–36. https://doi.org/10.1038/325031a0

Chan, Y. L., Schanzenbach, D., & Hickerson, M. J. (2014). Detecting concerted demographic response across community assemblages using hierarchical approximate Bayesian computation. *Molecular Biology and Evolution*, *31*, 2501–15. https://doi.org/10.1093/molbev/msu187

Condon, E., & Cukier, M. (2016). Using Approximate Bayesian Computation to Empirically Test Email Malware Propagation Models Relevant to Common Intervention Actions. In *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)* (pp. 287–297). IEEE. https://doi.org/10.1109/ISSRE.2016.24

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–8. https://doi.org/10.1016/j.tree.2010.04.001

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012

Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Molecular Biology and Evolution*, *15*, 1788–1790. https://doi.org/10.1093/oxfordjournals.molbev.a025905

Dussex, N., Wegmann, D., & Robertson, B. C. (2014). Postglacial expansion and not human influence best explains the population structure in the endangered kea (Nestor notabilis). *Molecular Ecology*, *23*, 2193–2209. https://doi.org/10.1111/mec.12729

Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*, 1332–1334. https://doi.org/10.1093/bioinformatics/btr124

Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*, 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, *8*, e79667.

Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, *30*, 1486–1487. https://doi.org/10.1093/bioinformatics/btu041

Guillemaud, T., Beaumont, M. A., Ciosi, M., Cornuet, J.-M., & Estoup, A. (2010). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, *104*, 88–99. https://doi.org/10.1038/hdy.2009.92

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695

Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS Genetics*, *9*, e1003521–e1003521. https://doi.org/10.1371/journal.pgen.1003521

Holliday, J. a, Yuen, M., Ritland, K., & Aitken, S. N. (2010). Postglacial history of a widespread conifer produces inverse clines in selective neutrality tests. *Molecular Ecology*, *19*, 3857–64. https://doi.org/10.1111/j.1365-294X.2010.04767.x

Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*, 1197–1206.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Leuenberger, C., & Wegmann, D. (2010). Bayesian computation and model selection without likelihoods. *Genetics*, *184*, 243–252. https://doi.org/10.1534/genetics.109.109058

Li, S., & Jakobsson, M. (2012). Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. *BMC Genetics*, *13*, 22–22. https://doi.org/10.1186/1471-2156-13-22

Li, Y., Stocks, M., Hemmila, S., Kallman, T., Zhu, H., Zhou, Y., … Lascoux, M. (2010). Demographic histories of four spruce (Picea) species of the Qinghai-Tibetan Plateau and neighboring areas inferred from multiple nuclear loci. *Molecular Biology and Evolution*, *27*, 1001–1014. https://doi.org/10.1093/molbev/msp301

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2016). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, syw077-syw077. https://doi.org/10.1093/sysbio/syw077

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., … Wang, J. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*, 785–794. https://doi.org/10.1016/j.cell.2014.03.054

Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the coalescent process. *Genetics*, *189*, 977–987. https://doi.org/10.1534/genetics.111.129569

Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, *367*, 1747–1757.

Marin, J. M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, *22*, 1167–1180. https://doi.org/10.1007/s11222-011-9288-2

Marin, J.-M., Pillai, N. S., Robert, C. P., & Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 833–859.

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013). Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genetics*, *9*, e1003942. https://doi.org/10.1371/journal.pgen.1003942

Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*, 2841–2847. https://doi.org/10.1111/mec.12350

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*, e37558. https://doi.org/10.1371/journal.pone.0037558

Pavlidis, P., Laurent, S., & Stephan, W. (2010). msABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Molecular Ecology Resources*, *10*, 723–727. https://doi.org/10.1111/j.1755-0998.2010.02832.x

Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., & French, N. P. (2013). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, *13*, 67–82. https://doi.org/10.1515/sagmb-2013-0012

Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., & Chikhi, L. (2012). Genetic data suggest a natural prehuman origin of open habitats in northern Madagascar and question the deforestation narrative in this region. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 13028–33. https://doi.org/10.1073/pnas.1200153109

R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* Retrieved from https://www.R-project.org/

Robert, C. P., Cornuet, J. M., Marin, J. M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 15112–15117. https://doi.org/10.1073/Pnas.1102900108

Robinson, J. D., Bunnefeld, L., Hearn, J., Stone, G. N., & Hickerson, M. J. (2014). ABC inference of multi-population divergence with admixture from unphased population genomic data. *Molecular Ecology*, *23*, 4458–4471. https://doi.org/10.1111/mec.12881

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*, 919–925. https://doi.org/10.1038/ng.3015

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, *16*, 727–740. https://doi.org/10.1038/nrg4005

Shafer, A. B. A., Gattepaille, L. M., Stewart, R. E. A., & Wolf, J. B. W. (2015). Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: In silico evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, *24*, 328–345. https://doi.org/10.1111/mec.13034

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, n/a-n/a. https://doi.org/10.1111/2041-210X.12700

Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., & Chikhi, L. (2012). Population divergence with or without admixture: selecting models using an ABC approach. *Heredity*, *108*, 521–530. https://doi.org/10.1038/hdy.2011.116

Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: efficiently simulating long sequences using the approximated coalescent with recombination. *Bioinformatics (Oxford, England)*, *31*, 1680–2. https://doi.org/10.1093/bioinformatics/btu861

Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice in Approximate Bayesian Computation. *PLoS ONE*, *9*, 1–13. https://doi.org/10.1371/journal.pone.0099581

Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate Bayesian Computation. *PLoS Computational Biology*, *9*. https://doi.org/10.1371/journal.pcbi.1002803

Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116–116. https://doi.org/10.1186/1471-2105-11-116

Weyant, A., Schafer, C., & Wood-Vasey, W. M. (2013). Likelihood-free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty. *The Astrophysical Journal*, *764*, 116. https://doi.org/10.1088/0004-637X/764/2/116

Wiuf, C., & Hein, J. (1999). Recombination as a Point Process along Sequences. *Theoretical Population Biology*, *55*, 248–259.

Zinck, J. W. R., & Rajora, O. P. (2016). Post-glacial phylogeography and evolution of a wide-ranging highly-exploited keystone forest tree, eastern white pine (Pinus strobus) in North America: single refugium, multiple routes. *BMC Evolutionary Biology*, *16*, 56–56. https://doi.org/10.1186/s12862-016-0624-1

## Data accessibility

All relevant information to reproduce this study is included in this manuscript and supporting information.

## Author contributions

J.S.E and S.N.A conceived the study. J.S.E performed simulations and analysed the data. J.S.E wrote the manuscript with input from from S.N.A.

## Supporting information

Additional supporting information including methods, figures and scripts can be found online.

# Figure and table captions

**Figure 1.** Demographic models. a) Model 1: A three-parameter model of expansion featuring colonization of new population 2 by 2 diploid individuals from population 1 at time $T_{EXP}$. Population 1 is of constant size $N_1$, whereas population 2 grows exponentially to size $N_2$, its size at present. b) Model 2: the number of founders of population 2 is a variable parameter. c) Model 3: a per-generation migration rate from population 1 to population 2 is added as a parameter. d) Model 4 includes all 5 parameters: $N_1$, $N_2$, $T_{EXP}$, $N_{02}$, and $m_{21}$.

**Figure 2.** Relative prediction error (RPE) calculated from the results of ABC analyses of 20 different combinations of demographic models and sampling designs (x-axis). For each combination, ABC was performed on simulated datasets summarized with statistics including linkage-based measures (hap. phase 1) and on the same set of simulations summarized with only SNP-based statistics (hap. phase 0). RPE values were calculated from the ABC estimation results of 1000 datasets with parameter values randomly drawn from their prior distributions.

**Figure 3.** Width of the 95% highest posterior density intervals calculated from the results of ABC analyses of 20 different combinations of demographic models and sampling designs. Error bars represent standard errors (N=100 PODs). See caption of figure 2 for more details.

**Figure 4.** Accuracy of parameter estimates for model 1. Within a plot, each datapoint corresponds to the estimated value of the parameter (mode of the posterior) vs. the true parameter value for one POD. Results are shown for a total of 100 PODs. Error bars correspond to the 95% HDI around the estimate. a) Results with datasets of type 1 (10,000 sequences of 100bp). Top panel shows results on unphased datasets, bottom panel shows results for phased datasets. b) Results with datasets of type 5 (100 sequences of 10,000bp). Top panel shows results on unphased datasets, bottom panel shows results for phased datasets.

**Figure 5.** RPE of model parameters for different fixed values of $T_{EXP}$. Results are shown for ABC runs with datasets of type 1 (10k sequences, 100-bp long). For a given parameter, results from different models are shown in the same plot window with different characters and colours. To see results for other model-dataset combinations as well as 95% HDI results, please see supporting information.

**Figure 6.** RPE and bootstrapped confidence intervals of model 2 parameters under different sequencing strategies and per-nucleotide error rates. N corresponds to the number of diploid individuals sequenced, dp to the mean individual sequencing depth. “perf” corresponds to perfect datasets whereas “err0”, “err0.001” and “err0.01” correspond to datasets where the sequencing process was simulated, with depth sampling and errors introduced at rates 0, 0.001, and 0.01 substitutions per nucleotide respectively. 13 PODs were used for each treatment.

**Figure 7.** RPE calculated from 100 datasets for models 1 to 4 using two different inference methods: ABC, computed on SNP-level summary statistics, and approximate composite likelihood, computed from the SFS. In both cases, datasets had 10,000 sequences of 100bp genotyped in 20 diploid individuals.

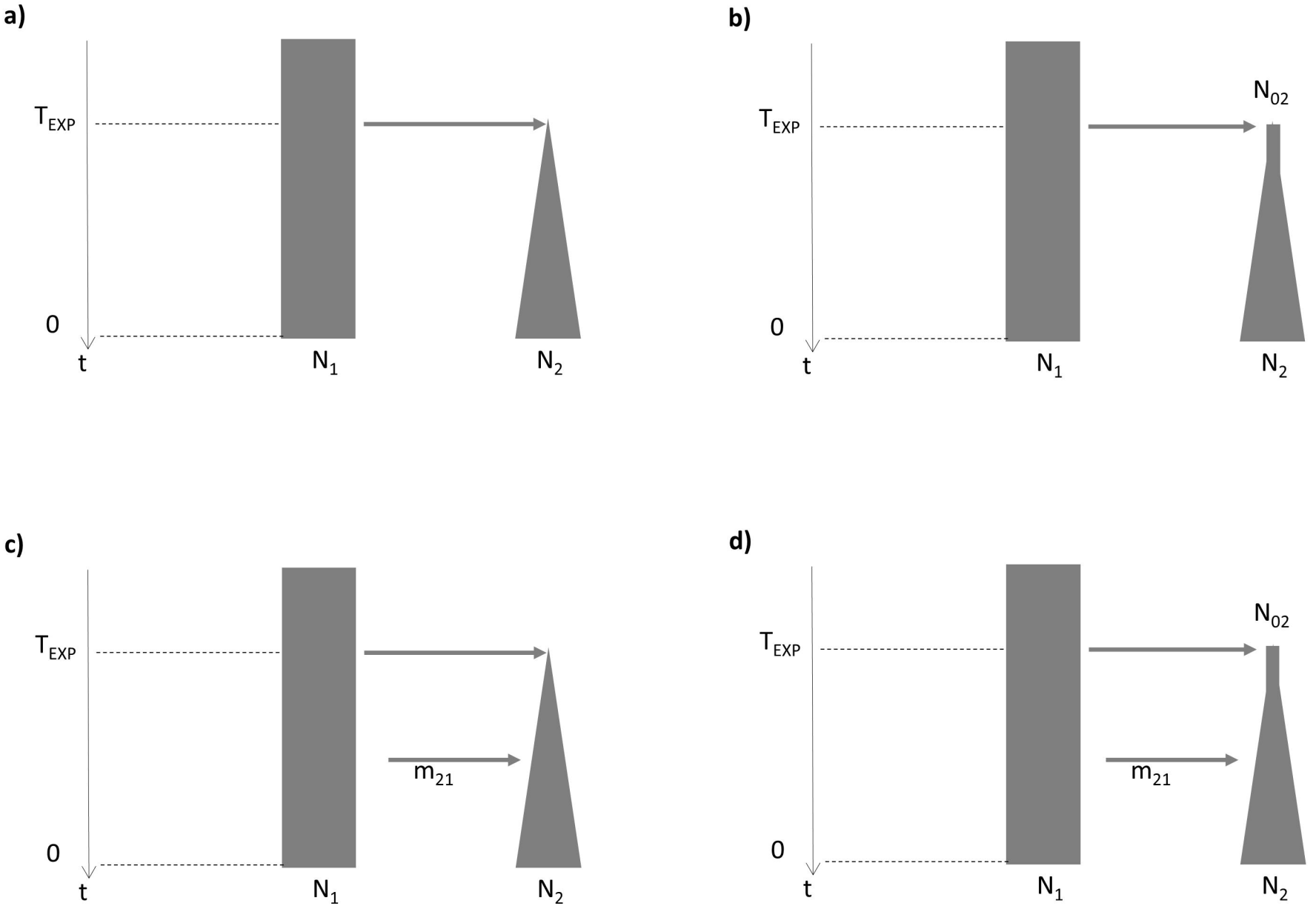
**Figure 8.** Width of the 95% HDI from ABC results, compared to 95% CI from the SFS inference method. For each of the four demographic models, the same 10 simulated datasets were used as pseudo-observed datasets for both the ABC and the SFS runs. HDI and CI widths were calculated from 100 bootstraps. Numbers correspond to the coverage of 95% CI (out of 10 PODs). PODs had 10,000 sequences of 100bp genotyped in 20 diploid individuals.

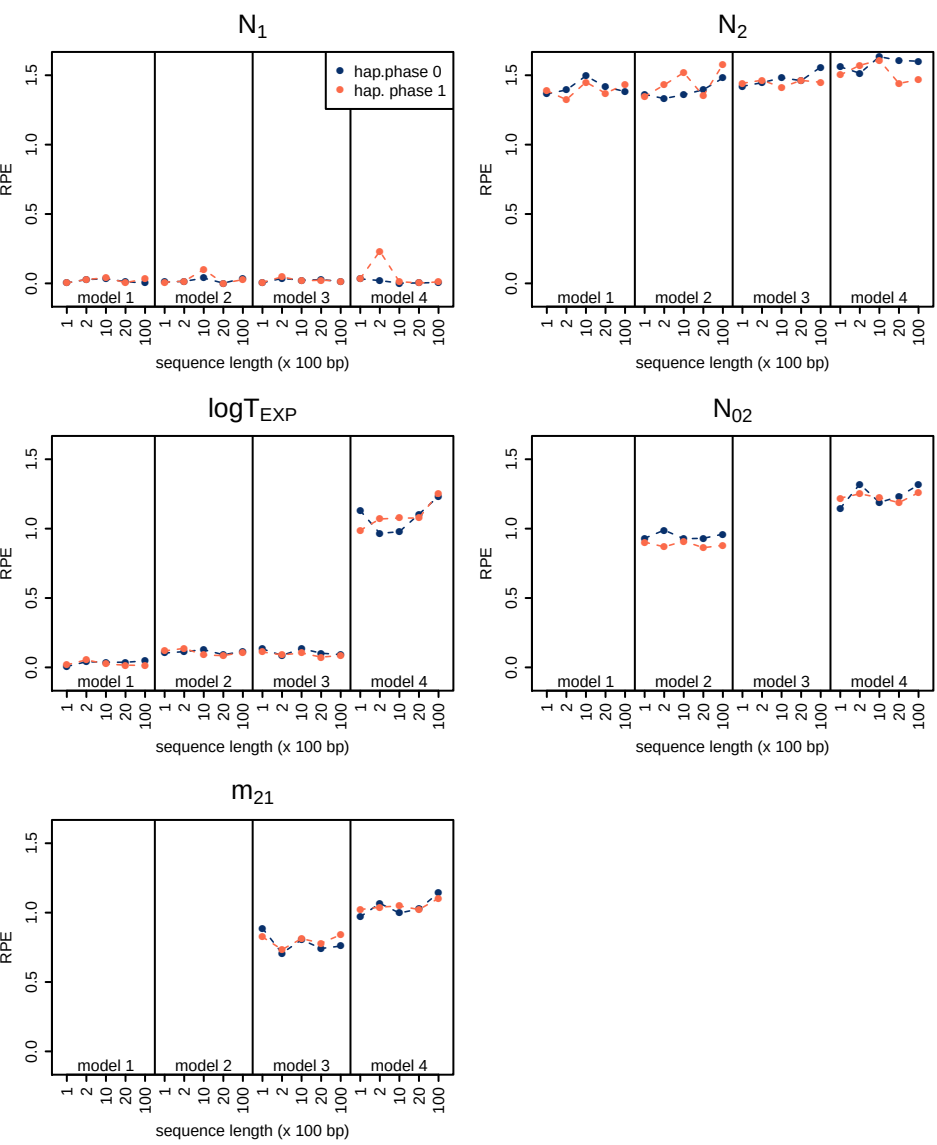
**Table 1.** Model parameters with their associated prior ranges

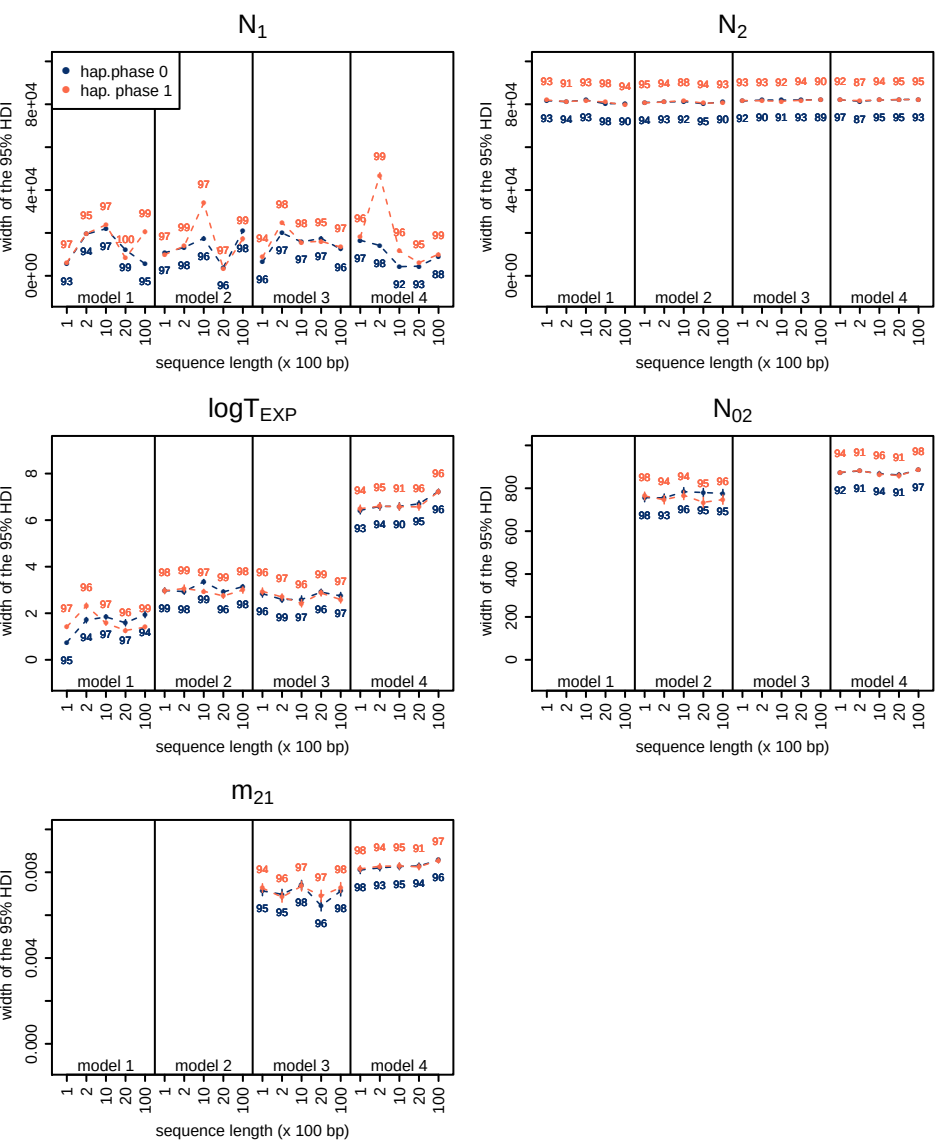
| estimated in models | Parameter | Symbol | Prior range | Unit |
|---|---|---|---|---|
| - | Mutation rate | $\mu$ | 0.000000009 | - |
| - | recombination rate | R | 0.00000001 | - |
| 1,2,3,4 | population size 1 | $N_1$ | U(10,000:100,000) | ind. |
| 1,2,3,4 | population size 2 | $N_2$ | U(10,000:100,000) | ind. |
| 1,2,3,4 | time of expansion | $T_{EXP}$ | logU(2:10,000) | gen |
| 2,4 | initial population size 2 | $N_{02}$ | U(2:1000) | ind. |
| 3,4 | migration rate from 1 to 2 | $m_{21}$ | U(0.001:0.01) | - |

**Table 2.** Description of the 5 types of simulated datasets

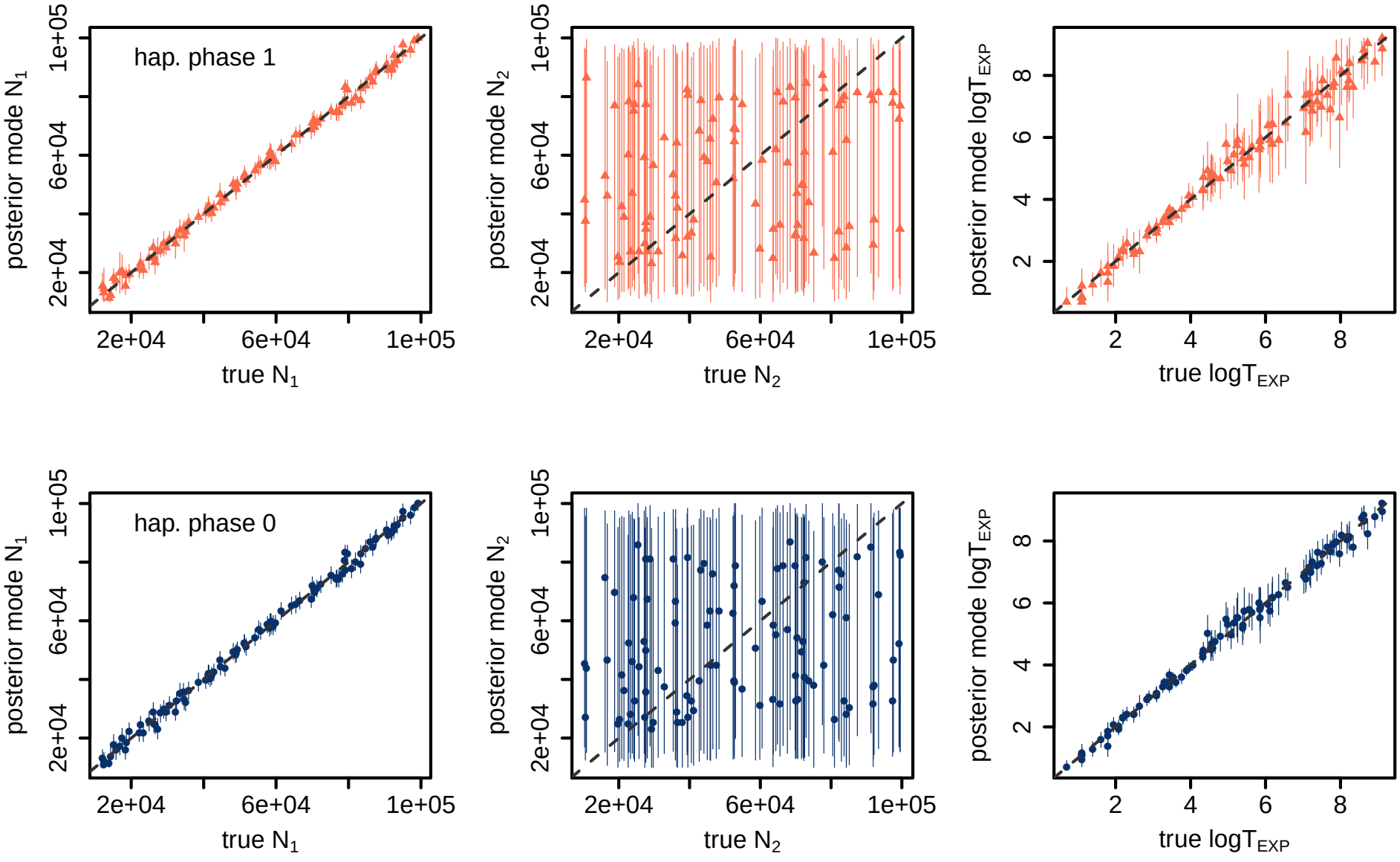
| | number of sequences | sequence length (bp) | number of diploid individuals |
|---|---|---|---|
| 1 | 10,000 | 100 | 20 |
| 2 | 5,000 | 200 | 20 |
| 3 | 1,000 | 1,000 | 20 |
| 4 | 500 | 2,000 | 20 |
| 5 | 100 | 10,000 | 20 |


a)
$T_{EXP}$
0
t
$N_1$
$N_2$
b)
$T_{EXP}$
$N_{02}$
0
t
$N_1$
$N_2$
c)
$T_{EXP}$
$m_{21}$
0
t
$N_1$
$N_2$
d)
$T_{EXP}$
$N_{02}$
$m_{21}$
0
t
$N_1$
$N_2$

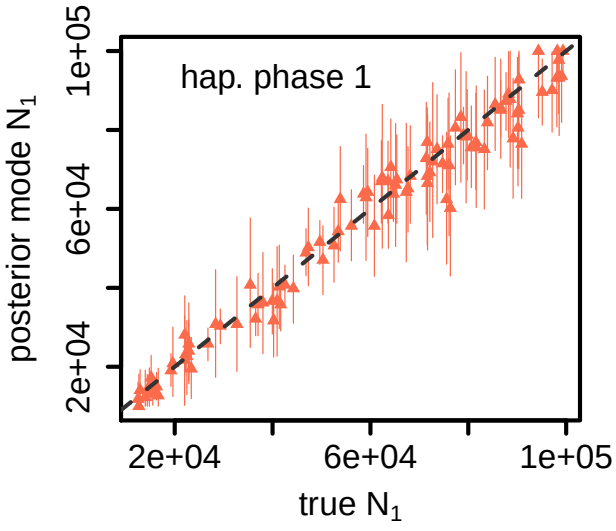

$N_1$
hap.phase 0
hap. phase 1
RPE
0.0
0.5
1.0
1.5
model 1
model 2
model 3
model 4
1
2
10
20
100
sequence length (x 100 bp)
$N_2$
$logT_{EXP}$
$N_{02}$
$m_{21}$

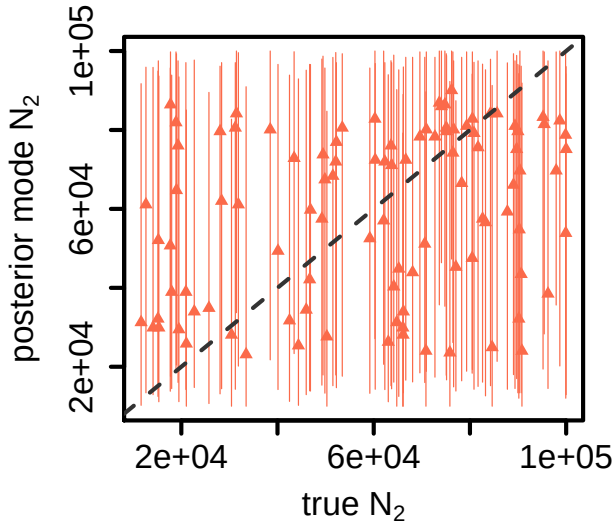

N1
hap.phase 0
hap. phase 1
width of the 95% HDI
0e+00
4e+04
8e+04
model 1
model 2
model 3
model 4
1
2
10
20
100
sequence length (x 100 bp)
N2
logTEXP
0
2
4
6
8
N02
0
200
400
600
800
m21
0.000
0.004
0.008

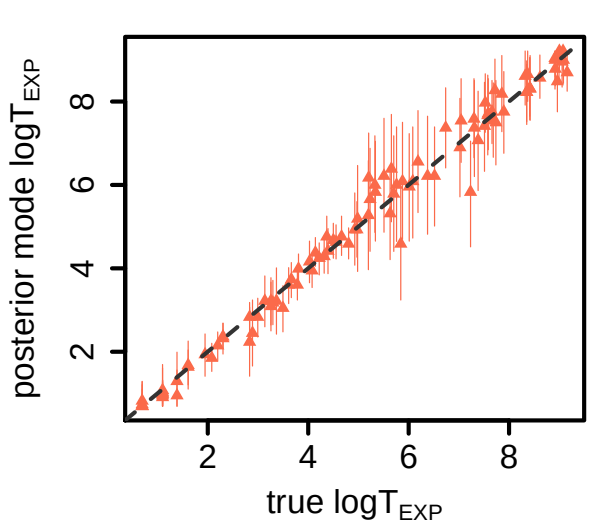
**a.**

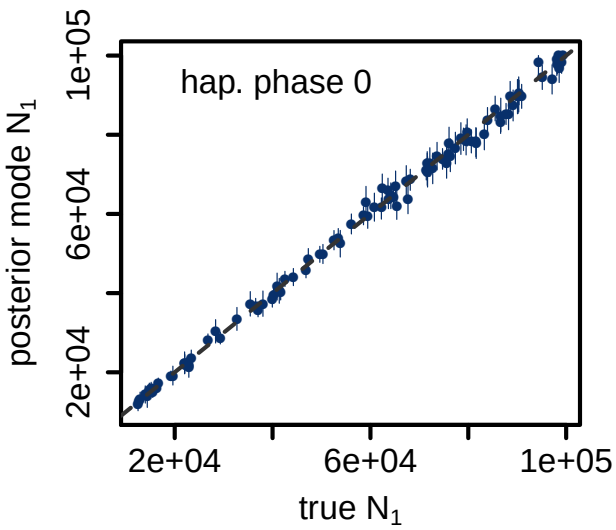

hap. phase 1
hap. phase 0
posterior mode $N_1$
posterior mode $N_2$
posterior mode $logT_{EXP}$
true $N_1$
true $N_2$
true $logT_{EXP}$
2e+04
6e+04
1e+05
2
4
6
8

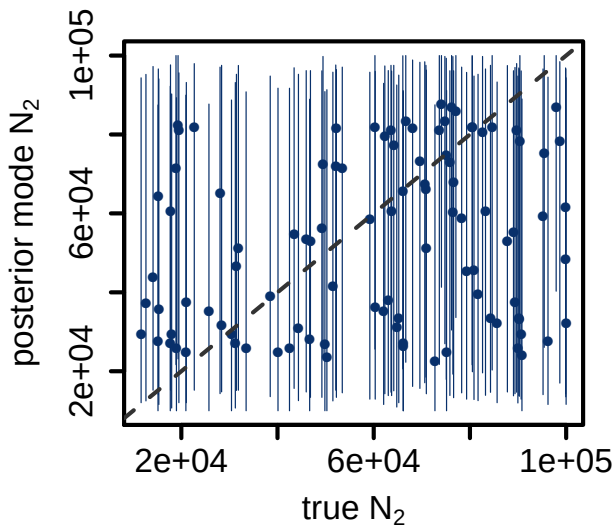
**b.**

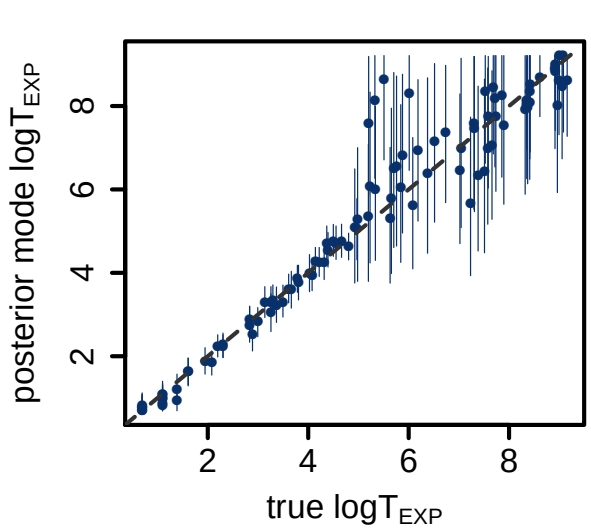

hap. phase 1
posterior mode $N_1$
true $N_1$
2e+04
6e+04
1e+05

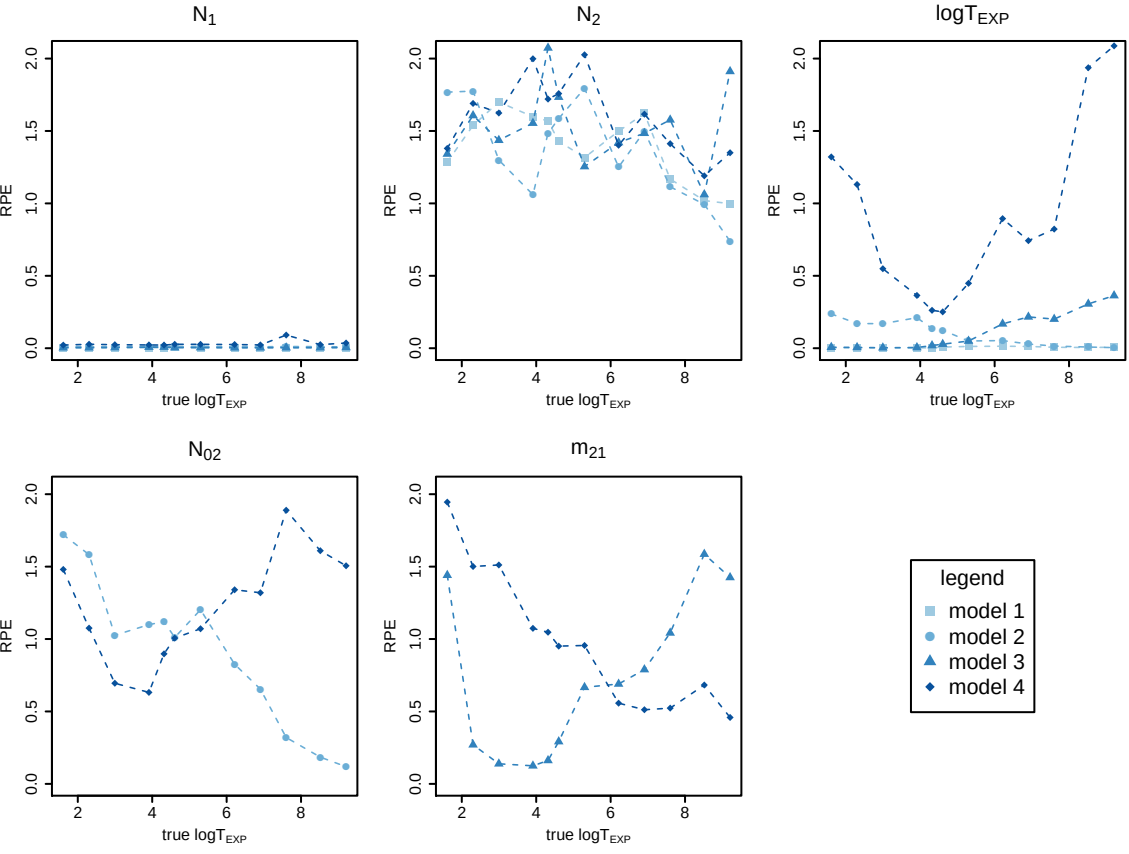

posterior mode $N_2$
true $N_2$
2e+04
6e+04
1e+05

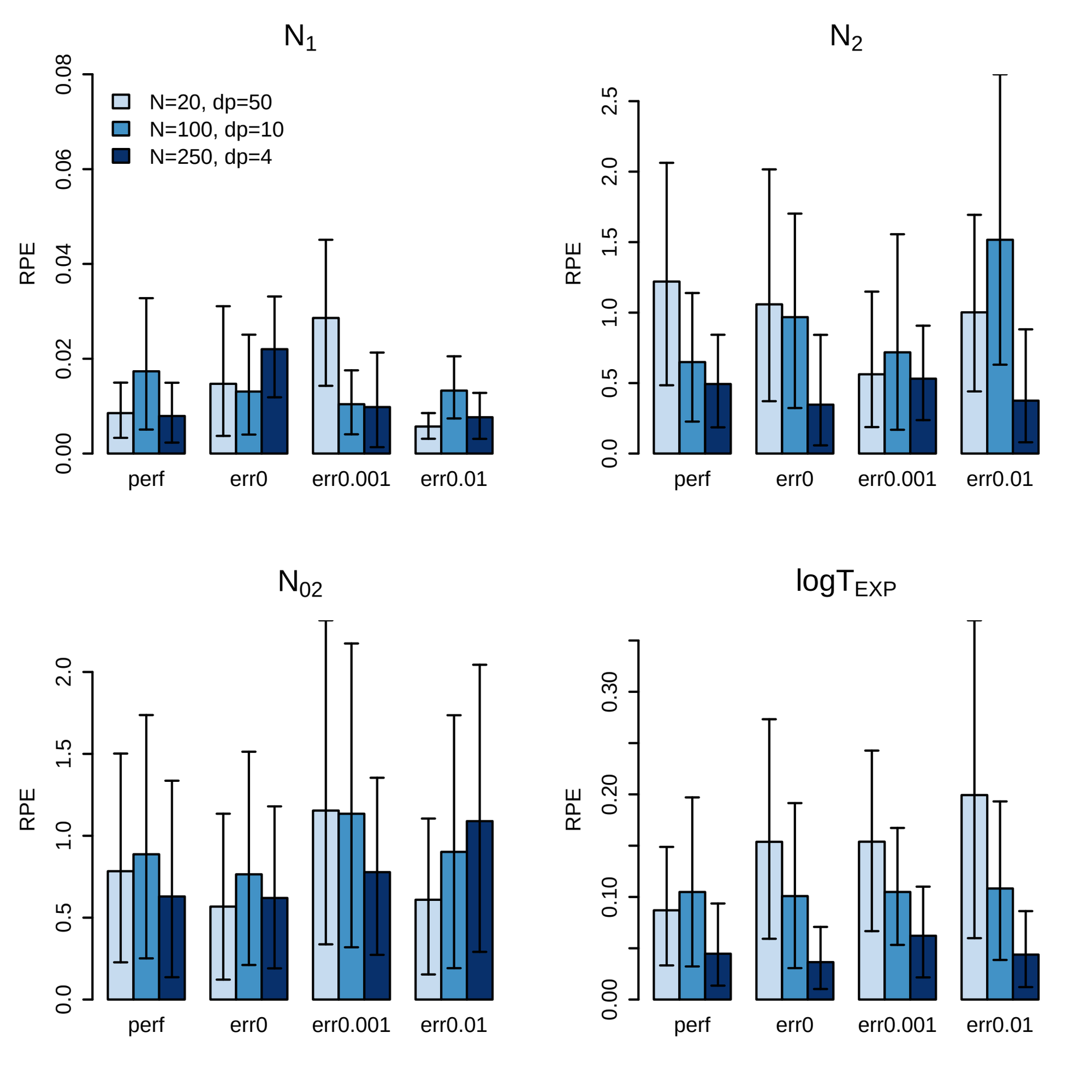

posterior mode $logT_{EXP}$
true $logT_{EXP}$
2
4
6
8

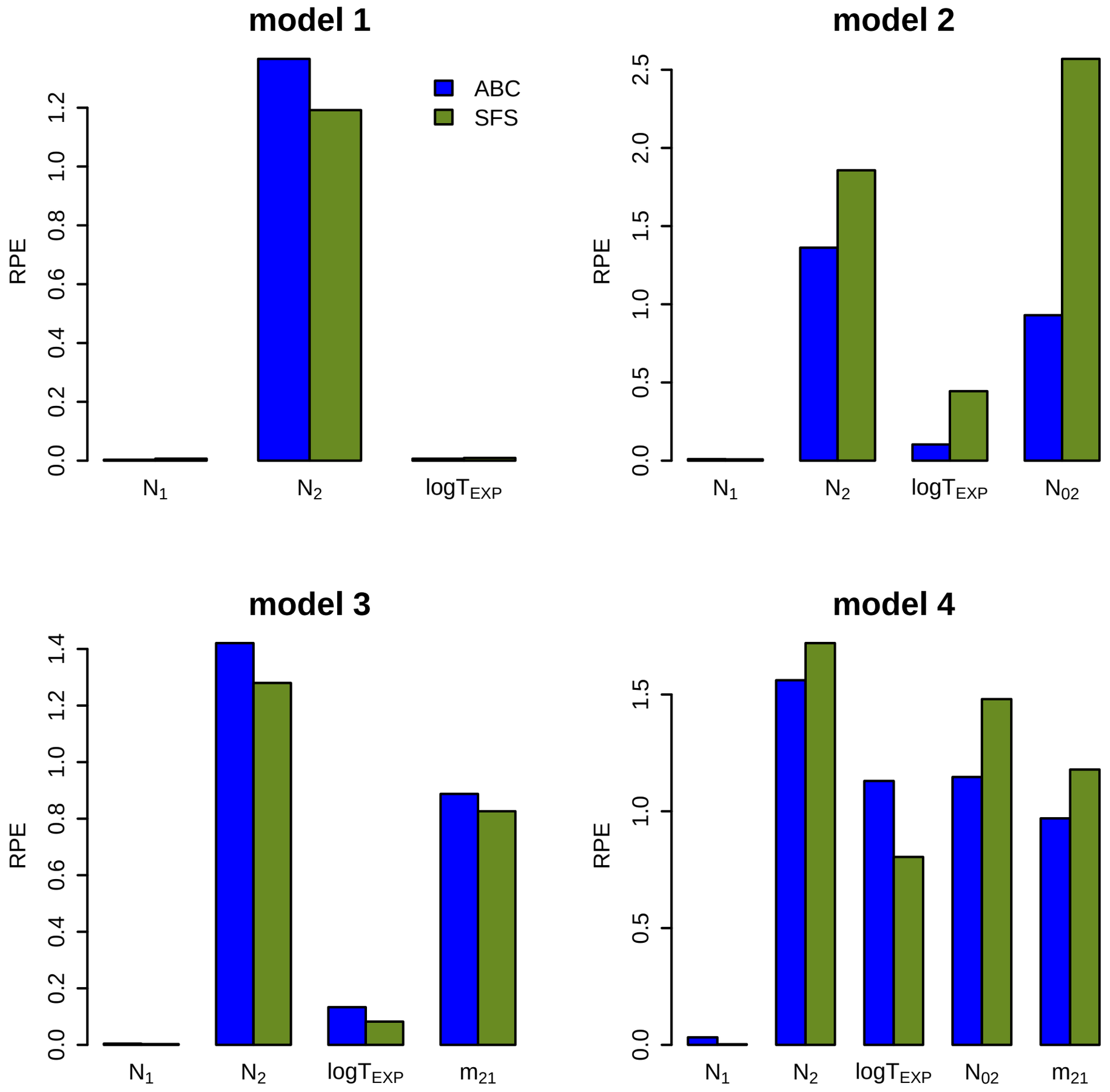

hap. phase 0
posterior mode $N_1$
true $N_1$
2e+04
6e+04
1e+05

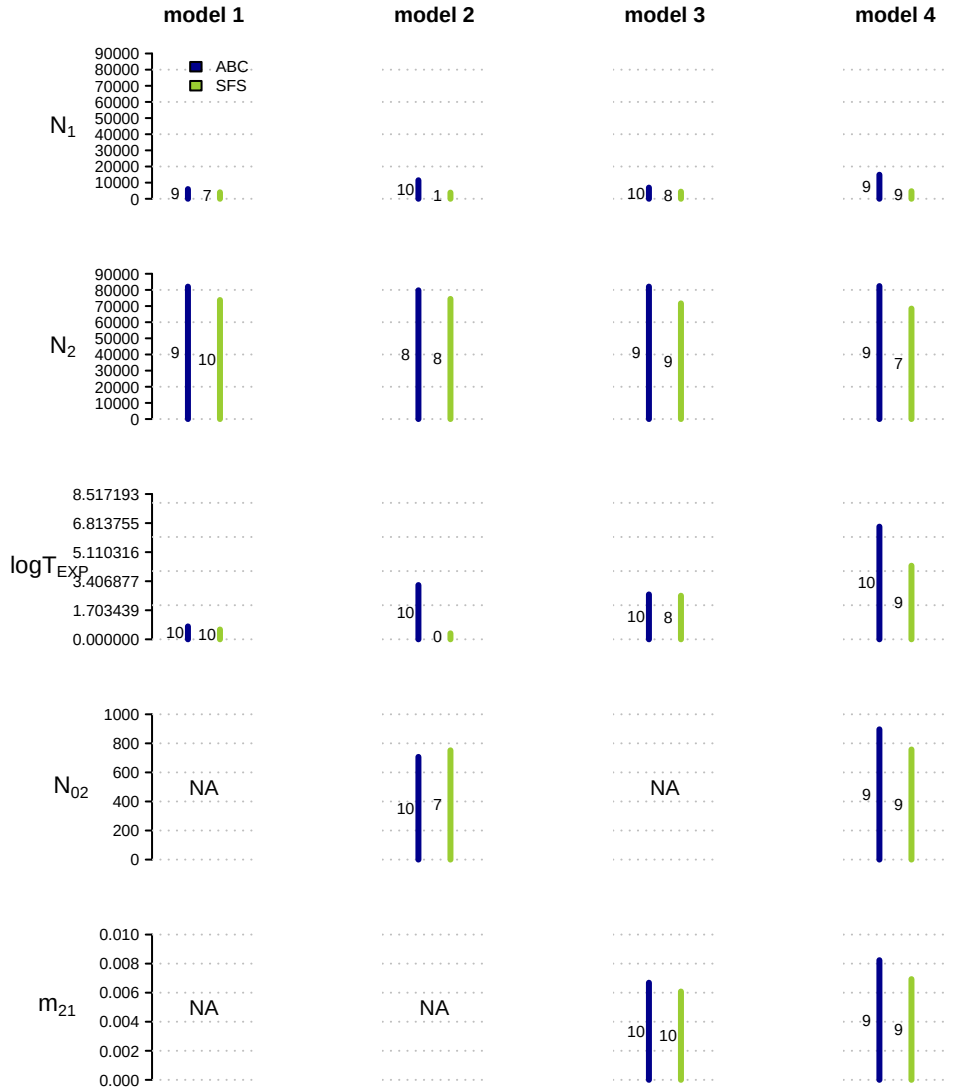

posterior mode $N_2$
true $N_2$
2e+04
6e+04
1e+05


posterior mode $logT_{EXP}$
true $logT_{EXP}$
2
4
6
8


N1
N2
logTEXP
N02
m21
RPE
true logTEXP
0.0
0.5
1.0
1.5
2.0
2
4
6
8
legend
model 1
model 2
model 3
model 4


$N_1$
$N_2$
$N_{02}$
$logT_{EXP}$
N=20, dp=50
N=100, dp=10
N=250, dp=4
RPE
perf
err0
err0.001
err0.01
0.00
0.02
0.04
0.06
0.08
0.0
0.5
1.0
1.5
2.0
2.5
0.10
0.20
0.30


model 1
ABC
SFS
RPE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
N1
N2
logTEXP
model 2
RPE
0.0
0.5
1.0
1.5
2.0
2.5
N1
N2
logTEXP
N02
model 3
RPE
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
N1
N2
logTEXP
m21
model 4
RPE
0.0
0.5
1.0
1.5
N1
N2
logTEXP
N02
m21


model 1
model 2
model 3
model 4
ABC
SFS
N1
N2
logTEXP
N02
m21
NA
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
8.517193
6.813755
5.110316
3.406877
1.703439
0.000000
1000
800
600
400
200
0.010
0.008
0.006
0.004
0.002
0.000