

1 **Title**

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

4

5 **Authors**

6 Joane S. Elleouet, Sally N. Aitken

7

8 **Address**

9 Department of Forest and Conservation Sciences, Faculty of Forestry, University of British Columbia,
10 3041-2424 Main Mall, Vancouver BC V6T 1Z4, Canada

11

12 **Keywords**

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

15

16 **Corresponding author**

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

18

19 **Running title**

20 Inferring recent population history with ABC

21 **Abstract**

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

39

40 **Introduction**

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

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

54

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

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

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

86

87 **Approximate Bayesian Computation and other approaches**

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

90 approach based on approximation of the composite likelihood (Excoffier & Foll, 2011). We chose to
91 explore ABC because of its versatility: It accommodates a wide spectrum of demographic models and
92 dataset types. Although it was originally developed for inferences in evolutionary biology, the statistical
93 framework of ABC has been extended to a variety of disciplines, from cell biochemistry and
94 epidemiology to neural networks, extending beyond the realm of biology into meteorology, astrophysics
95 (Weyant et al., 2013) and computer sciences (Condon & Cukier, 2016). ABC has been reviewed in a
96 number of publications and its algorithms and techniques are being refined constantly (Bertorelle et al.,
97 2010; Csilléry et al., 2010; Lintusaari et al., 2016; Marin et al., 2012; Sunnaker et al., 2013). For
98 applications in demographic inference using genetic data, the general ABC method involves the
99 following steps. First, a large number of datasets are simulated under a specific demographic model
100 using the coalescent (Kingman, 1982). Parameters used for simulations are drawn from prior
101 distributions that are pre-defined by the user. The simulated datasets are then compared to the
102 observed dataset through calculation of summary statistics. Finally, simulated datasets with the closest
103 vector of statistics to the vector of observed summary statistics are selected. A regression adjustment
104 based on the local relationship between statistics and parameters is then usually performed to
105 approximate the posterior distribution of each model parameter from the parameter values of selected
106 simulations. ABC is suitable when inferring models for which the likelihood function is intractable, as it
107 relies on approximating the likelihood function using a large number of simulations. However, each one
108 of the numerous steps in the implementation of ABC requires users to make empirical decisions. There is
109 particularly a need to improve our understanding of the relationship between the type of markers
110 obtained to build genetic datasets and the way genetic data is subsequently summarized on its power to
111 tease apart demographic models and produce accurate parameter estimates.

112

113 **Previous work exploring ABC**

114 The need to test the inference power of datasets for demographic models of interest has been
115 recognized in recent years, both in terms of model selection and parameter estimation. Robert et al.
116 (2011) warned against the use of insufficient summary statistics in ABC model choice, opening the door
117 to improved methods for model testing and the associated choice of summary statistics (Marin et al.,
118 2014; Prangle et al., 2013). Among theoretical results and general guidelines, Marin et al. (2014)
119 suggested the use of different sets of summary statistics for estimation and model selection. Several
120 studies show the use of preliminary simulations testing parameter estimation and model choice with
121 different number and length of markers and number of individuals (Sousa et al., 2012; Stocks et al.,
122 2014), type of molecular markers (Cabrera & Palsbøll, 2017) and choice of summary statistics and
123 models considered (Benazzo et al., 2015; Guillemaud et al., 2010; Li & Jakobsson, 2012; Sousa et al.,
124 2012; Stocks et al., 2014). As most scientists have switched to using genome-wide data, there is a need
125 to expand this set of simulation studies to test and understand the power of different types of genomic
126 data. As part of such an effort, Li & Jakobsson (2012) simulated large, phased genomic datasets
127 comparable to human genomic datasets at the time. Under 2-population split models, they found that
128 ABC produces accurate estimates for most but not all parameters and concluded ABC is well suited to
129 large genomic datasets summarized with LD-based statistics. Robinson et al. (2014) tested the effects of
130 the number and length of unphased genomic sequences and compared them to the effect of the
131 number of individuals sequenced for the inference of three-population admixture models. They found
132 that increasing the number and length of sequences was more beneficial than increasing sample size.
133 Shafer et al. (2015) investigated the power of ABC on short diploid sequences obtained by GBS. They
134 focused on a wide range of simple 1-population and 2-population models with bottleneck, growth,
135 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 $r = \log\left(\frac{N_{02}}{N_2}\right) / T_{EXP}$. 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 θ estimates) and summaries of the SFS (Tajima's D and Fay and Wu's H). These statistics were calculated

204 for each population and for the whole sample. The available statistics also include summaries of the 2d-
205 SFS: differentiation measures such as the pairwise F_{ST} and the number of private and shared
206 polymorphisms. Finally the Thomson estimator of T_{MRCA} and its variance were calculated for each
207 population and for the whole sample. To test the effect that knowing haplotype information has on
208 inference, the ABC analysis was performed twice on each model-dataset type combination. The first
209 time, we summarized data using only the statistics mentioned above, which are calculated at the SNP
210 level and therefore are available when the gametic phase of the diploid sequences is unknown. The
211 second inference was performed on the same dataset, but additional statistics Z_{ns} (Kelly, 1997), dv_k and
212 dv_h (Depaulis & Veuille, 1998) based on linkage information were used to summarize the data. These
213 additional statistics are calculated at the haplotype level and so are only available in cases where the
214 gametic phase of the diploid sequences is known. For each set of simulations, we computed the mean
215 and variance of every statistic over all sequence markers in the dataset. As a result, 58 statistics were
216 computed for datasets with known gametic phases (hereafter referred to as “phased”, or “hap.phase
217 1”), and 43 statistics were computed for datasets with unknown gametic phases (hereafter referred to as
218 “unphased”, or “hap.phase 0”).

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

223

224 **Pseudo-observed datasets**

225 For each set of 1M simulations, we created a corresponding set of 100 pseudo observed
226 datasets (PODs), with parameters randomly chosen from the same priors as for the set of 1M
227 simulations. By doing so we assume that priors are reliable and reflect the true, unknown distribution of
228 the PODs. These were then summarized with the same summary statistics as their corresponding set of
229 1M simulations.

230

231 **ABC estimation**

232 We performed the ABC estimation using each POD as the observed dataset to obtain parameter
233 estimates. The standard ESTIMATE algorithm from the program ABCtoolbox (Wegmann et al., 2010) was
234 used for all ABC computations to create posterior probabilities from the corresponding set of 1M
235 simulations, with a post-sampling regression adjustment through ABC-GLM (Leuenberger & Wegmann,
236 2010). We fixed the tolerance parameter to 10^{-3} , a compromise between having a tolerance threshold
237 value as low as possible (Li & Jakobsson, 2012) and keeping an appropriate number of simulations to
238 estimate the posterior from.

239

240 **Validation**

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

$$244 \quad (2) \mathcal{E} = \frac{\sum_{j=1}^{j=i} (\widehat{\theta}_j - \theta_j^*)^2}{Var(\theta)} \times \frac{1}{i}$$

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

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

256

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

258 To test the effect of the time of expansion on the precision of the ABC estimation, we created
259 100 PODs for each set of 1M simulations and 12 fixed values of $\log T_{EXP}$ spanning the prior range. RPE and
260 95%HDI were calculated from the results of each set of 100 PODs.

261

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

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

266 sequencing depth, which in turn can affect variant calling and estimation of allelic frequencies
267 (Fumagalli, 2013). We explored this challenge focusing on model 2 and dataset type 2. We chose a
268 realistic fixed sequencing effort and derived 3 fixed sampling strategies from it: 250 sampled individuals
269 at a mean individual depth of 4, 100 individuals with depth 10, and 20 individuals with depth 50. We
270 then incorporated three per-nucleotide sequencing error rates (0 , 10^{-2} , 10^{-3}), and applied them to each
271 category described above. The resulting 9 categories of PODs, as well as “perfect” datasets (no depth
272 sampling and no error) were all simulated using the same 10 parameter combinations. Further details
273 about the creation of “imperfect” PODs can be found in the supplemental methods. Once these
274 imperfect PODs were created and summarized, ABC was performed to estimate their true parameter
275 values. Two additional sets of 1M simulations needed to be created to match the number of individuals
276 sampled per population: one with 100 diploids per population, and the second with 250. In the latter
277 case, we only created 610,000 simulations because of computation time limitations. The same tolerance
278 (0.001) as all other runs was used for the estimation.

279

280 **Comparing ABC and SFS estimation**

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

286

287 **Results**

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

293

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

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

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

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

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

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

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

325

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

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

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

333

334 **Quality of parameter estimates across prior ranges**

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

347

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

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

352 and S3). The precision of inference on model 1 is higher at low T_{EXP} values and decreases at $\log T_{EXP} > 4$.
353 Conversely, for model 2, older events are generally better inferred: estimates of T_{EXP} and N_{02} increase in
354 precision as T_{EXP} increases, as shown by the RPE (fig. S2, p.2) and the 95% HDI (fig. S3, p.2). Model 3
355 shows the best results for moderately recent expansion events ($3 < \log T_{EXP} < 4$), as shown by RPE and
356 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
357 for all parameters, with RPE values mostly above 0.5.

358

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

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

371

372 **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.

417 One needs to be aware of the difficulties associated with the use of LD information. Firstly, ABC
418 on phased data requires reasonable knowledge of recombination rates and variability across the
419 genome. The recombination rate needs to be included as a parameter along with demographic
420 parameters, or as a nuisance parameter with a hyper-prior. Secondly, simulating the coalescent with
421 recombination is a complicated process and comes at high computational costs (McVean & Cardin,
422 2005). With high recombination rates or very long sequences, coalescent simulations might take so long
423 to run that one would instead use a more efficient inference method than ABC. Moreover, translating
424 genome-wide observed data into a set of summary statistic values that are readily useable by ABC
425 programs and comparable to simulated datasets can be a challenge. File input formats in most programs
426 are currently not compatible with sequence information, and many summary statistics programs do not
427 offer haplotype-level calculations. Thirdly, when aligning reads to a fragmented and incomplete
428 reference genome, as is often the case for non-model organisms, defining haplotypes can be tricky. One
429 also needs to address problems of sequencing errors, paralogous sequences and imperfect mapping.
430 Inevitable sequencing uncertainties will affect haplotype statistics more strongly than single-SNP
431 diversity measures. Data processing errors and filters can severely bias inferences, to the extent of
432 supporting the wrong demographic model, as revealed by Shafer et al. (2016). Finally, targeted sequence
433 capture will result in thousands of markers of various lengths. Setting up simulations that correspond
434 closely to an observed dataset requires approximating the distribution of sequence lengths, and this may
435 also affect inferences, especially if variances of summary statistics are included at the data
436 summarization step. Considering the difficulty of obtaining reliable haplotype information in non-model
437 organisms, the potential difficulties of adapting the use of long sequences to currently available ABC
438 programs, and computational time, our results tend to suggest that using SNP-level information from
439 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 rigorous 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,

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

488 ABC has proven moderately useful for demographic inference with long, genome-wide
489 haplotypes but comparisons with alternative approaches are scarce. Notable examples include
490 Nadachowska-Brzyska *et al.* (2013), who used ABC and PSMC in a complementary way. Robinson *et al.*
491 (2014) compared their ABC results with an exact likelihood method developed by Lohse *et al.* (2011) and
492 found that ABC resulted in more uncertainty, especially in model comparisons. As ABC performance with
493 linkage information needs to be further explored, comparisons to emerging analytical methods based on
494 whole genomes or long sequences such as MSMC (Schiffels & Durbin, 2014) or identity-by-descent
495 haplotype sharing (Harris & Nielsen, 2013) will greatly help refine methods for demographic inference
496 using data at a genomic scale.

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

502

503 **Acknowledgements**

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

508

509

510 **References**

- 511 Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored
512 populations. *Genetics*, *164*, 1139–1160.
- 513 Benazzo, A., Ghirotto, S., Vilaça, S. T., & Hoban, S. (2015). Using ABC and microsatellite data to detect
514 multiple introductions of invasive species from a single source. *Heredity*, *115*, 262–272.
515 <https://doi.org/10.1038/hdy.2015.38>
- 516 Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over
517 space and time: Some cons, many pros. *Molecular Ecology*, *19*, 2609–2625.
518 <https://doi.org/10.1111/j.1365-294X.2010.04690.x>
- 519 Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension
520 reduction methods in approximate Bayesian computation. *Statistical Science*, *28*, 189–208.
521 <https://doi.org/10.1214/12-STS406>
- 522 Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from
523 large samples of genome-wide molecular data: an approximate Bayesian computation approach.
524 *PLOS Genetics*, *12*, e1005877–e1005877. <https://doi.org/10.1371/journal.pgen.1005877>
- 525 Cabrera, A. A., & Palsbøll, P. J. (2017). Inferring past demographic changes from contemporary genetic
526 data: A simulation-based evaluation of the ABC methods implemented in DIYABC. *Molecular
527 Ecology Resources*. <https://doi.org/10.1111/1755-0998.12696>
- 528 Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*,
529 31–36. <https://doi.org/10.1038/325031a0>
- 530 Chan, Y. L., Schanzenbach, D., & Hickerson, M. J. (2014). Detecting concerted demographic response
531 across community assemblages using hierarchical approximate Bayesian computation. *Molecular
532 Biology and Evolution*, *31*, 2501–15. <https://doi.org/10.1093/molbev/msu187>
- 533 Condon, E., & Cukier, M. (2016). Using Approximate Bayesian Computation to Empirically Test Email
534 Malware Propagation Models Relevant to Common Intervention Actions. In *2016 IEEE 27th
535 International Symposium on Software Reliability Engineering (ISSRE)* (pp. 287–297). IEEE.
536 <https://doi.org/10.1109/ISSRE.2016.24>
- 537 Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation
538 (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–8.
539 <https://doi.org/10.1016/j.tree.2010.04.001>
- 540 Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-
541 wide genetic marker discovery and genotyping using next-generation sequencing. *Nature
542 Reviews Genetics*, *12*, 499–510. <https://doi.org/10.1038/nrg3012>
- 543 Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an
544 infinite-site model. *Molecular Biology and Evolution*, *15*, 1788–1790.
545 <https://doi.org/10.1093/oxfordjournals.molbev.a025905>
- 546 Dussex, N., Wegmann, D., & Robertson, B. C. (2014). Postglacial expansion and not human influence best
547 explains the population structure in the endangered kea (*Nestor notabilis*). *Molecular Ecology*,
548 *23*, 2193–2209. <https://doi.org/10.1111/mec.12729>
- 549 Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity
550 under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*, 1332–1334.
551 <https://doi.org/10.1093/bioinformatics/btr124>
- 552 Fearnhead, P., & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian
553 computation: semi-automatic approximate Bayesian computation. *Journal of the Royal
554 Statistical Society: Series B (Statistical Methodology)*, *74*, 419–474.
555 <https://doi.org/10.1111/j.1467-9868.2011.01010.x>

- 556 Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics
557 inferences. *PLoS One*, *8*, e79667.
- 558 Fumagalli, M., Vieira, F. G., Linderoth, T., & Nielsen, R. (2014). ngsTools: methods for population genetics
559 analyses from next-generation sequencing data. *Bioinformatics*, *30*, 1486–1487.
560 <https://doi.org/10.1093/bioinformatics/btu041>
- 561 Guillemaud, T., Beaumont, M. A., Ciosi, M., Cornuet, J.-M., & Estoup, A. (2010). Inferring introduction
562 routes of invasive species using approximate Bayesian computation on microsatellite data.
563 *Heredity*, *104*, 88–99. <https://doi.org/10.1038/hdy.2009.92>
- 564 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint
565 Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS*
566 *Genetics*, *5*, e1000695. <https://doi.org/10.1371/journal.pgen.1000695>
- 567 Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype
568 lengths. *PLoS Genetics*, *9*, e1003521–e1003521. <https://doi.org/10.1371/journal.pgen.1003521>
- 569 Holliday, J. a, Yuen, M., Ritland, K., & Aitken, S. N. (2010). Postglacial history of a widespread conifer
570 produces inverse clines in selective neutrality tests. *Molecular Ecology*, *19*, 3857–64.
571 <https://doi.org/10.1111/j.1365-294X.2010.04767.x>
- 572 Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*, 1197–1206.
- 573 Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248.
574 [https://doi.org/10.1016/0304-4149\(82\)90011-4](https://doi.org/10.1016/0304-4149(82)90011-4)
- 575 Leuenberger, C., & Wegmann, D. (2010). Bayesian computation and model selection without likelihoods.
576 *Genetics*, *184*, 243–252. <https://doi.org/10.1534/genetics.109.109058>
- 577 Li, S., & Jakobsson, M. (2012). Estimating demographic parameters from large-scale population genomic
578 data using Approximate Bayesian Computation. *BMC Genetics*, *13*, 22–22.
579 <https://doi.org/10.1186/1471-2156-13-22>
- 580 Li, Y., Stocks, M., Hemmilla, S., Kallman, T., Zhu, H., Zhou, Y., ... Lascoux, M. (2010). Demographic histories
581 of four spruce (*Picea*) species of the Qinghai-Tibetan Plateau and neighboring areas inferred
582 from multiple nuclear loci. *Molecular Biology and Evolution*, *27*, 1001–1014.
583 <https://doi.org/10.1093/molbev/msp301>
- 584 Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2016). Fundamentals and recent
585 developments in approximate Bayesian computation. *Systematic Biology*, *syw077-syw077*.
586 <https://doi.org/10.1093/sysbio/syw077>
- 587 Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., ... Wang, J. (2014). Population genomics
588 reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*, 785–794.
589 <https://doi.org/10.1016/j.cell.2014.03.054>
- 590 Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the
591 coalescent process. *Genetics*, *189*, 977–987. <https://doi.org/10.1534/genetics.111.129569>
- 592 Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden
593 of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, *367*,
594 1747–1757.
- 595 Marin, J. M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational
596 methods. *Statistics and Computing*, *22*, 1167–1180. <https://doi.org/10.1007/s11222-011-9288-2>
- 597 Marin, J.-M., Pillai, N. S., Robert, C. P., & Rousseau, J. (2014). Relevant statistics for Bayesian model
598 choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 833–859.
- 599 McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical*
600 *Transactions of the Royal Society B: Biological Sciences*, *360*, 1387–1393.
601 <https://doi.org/10.1098/rstb.2005.1673>

- 602 Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013).
603 Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-
604 genome re-sequencing data. *PLoS Genetics*, *9*, e1003942.
605 <https://doi.org/10.1371/journal.pgen.1003942>
- 606 Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-
607 sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*, 2841–2847.
608 <https://doi.org/10.1111/mec.12350>
- 609 Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and
610 Sample Allele Frequency Estimation from New-Generation Sequencing Data. *PLoS ONE*, *7*,
611 e37558. <https://doi.org/10.1371/journal.pone.0037558>
- 612 Pavlidis, P., Laurent, S., & Stephan, W. (2010). msABC: a modification of Hudson's ms to facilitate multi-
613 locus ABC analysis. *Molecular Ecology Resources*, *10*, 723–727. <https://doi.org/10.1111/j.1755-0998.2010.02832.x>
- 615 Prangle, D., Fearnhead, P., Cox, M. P., Biggs, P. J., & French, N. P. (2013). Semi-automatic selection of
616 summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular
617 Biology*, *13*, 67–82. <https://doi.org/10.1515/sagmb-2013-0012>
- 618 Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., & Chikhi, L. (2012). Genetic data suggest a natural
619 prehuman origin of open habitats in northern Madagascar and question the deforestation
620 narrative in this region. *Proceedings of the National Academy of Sciences of the United States of
621 America*, *109*, 13028–33. <https://doi.org/10.1073/pnas.1200153109>
- 622 R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for
623 Statistical Computing, Vienna, Austria*. Retrieved from <https://www.R-project.org/>
- 624 Robert, C. P., Cornuet, J. M., Marin, J. M., & Pillai, N. S. (2011). Lack of confidence in approximate
625 Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the
626 United States of America*, *108*, 15112–15117. <https://doi.org/10.1073/Pnas.1102900108>
- 627 Robinson, J. D., Bunnefeld, L., Hearn, J., Stone, G. N., & Hickerson, M. J. (2014). ABC inference of multi-
628 population divergence with admixture from unphased population genomic data. *Molecular
629 Ecology*, *23*, 4458–4471. <https://doi.org/10.1111/mec.12881>
- 630 Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple
631 genome sequences. *Nature Genetics*, *46*, 919–925. <https://doi.org/10.1038/ng.3015>
- 632 Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history.
633 *Nature Reviews Genetics*, *16*, 727–740. <https://doi.org/10.1038/nrg4005>
- 634 Shafer, A. B. A., Gattepaille, L. M., Stewart, R. E. A., & Wolf, J. B. W. (2015). Demographic inferences
635 using short-read genomic data in an approximate Bayesian computation framework: In silico
636 evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, *24*, 328–
637 345. <https://doi.org/10.1111/mec.13034>
- 638 Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016).
639 Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic
640 inference. *Methods in Ecology and Evolution*, n/a-n/a. <https://doi.org/10.1111/2041-210X.12700>
- 641 Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., & Chikhi, L. (2012). Population divergence
642 with or without admixture: selecting models using an ABC approach. *Heredity*, *108*, 521–530.
643 <https://doi.org/10.1038/hdy.2011.116>
- 644 Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: efficiently simulating long sequences using
645 the approximated coalescent with recombination. *Bioinformatics (Oxford, England)*, *31*, 1680–2.
646 <https://doi.org/10.1093/bioinformatics/btu861>

- 647 Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice
648 in Approximate Bayesian Computation. *PLoS ONE*, *9*, 1–13.
649 <https://doi.org/10.1371/journal.pone.0099581>
- 650 Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate
651 Bayesian Computation. *PLoS Computational Biology*, *9*.
652 <https://doi.org/10.1371/journal.pcbi.1002803>
- 653 Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: a versatile
654 toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116–116.
655 <https://doi.org/10.1186/1471-2105-11-116>
- 656 Weyant, A., Schafer, C., & Wood-Vasey, W. M. (2013). Likelihood-free Cosmological Inference with Type
657 Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty.
658 *The Astrophysical Journal*, *764*, 116. <https://doi.org/10.1088/0004-637X/764/2/116>
- 659 Wiuf, C., & Hein, J. (1999). Recombination as a Point Process along Sequences. *Theoretical Population*
660 *Biology*, *55*, 248–259.
- 661 Zinck, J. W. R., & Rajora, O. P. (2016). Post-glacial phylogeography and evolution of a wide-ranging
662 highly-exploited keystone forest tree, eastern white pine (*Pinus strobus*) in North America: single
663 refugium, multiple routes. *BMC Evolutionary Biology*, *16*, 56–56.
664 <https://doi.org/10.1186/s12862-016-0624-1>
665
666

667 **Data accessibility**

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

670

671

672 **Author contributions**

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

675

676 **Supporting information**

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

678

679 **Figure and table captions**

680
681 **Figure 1.** Demographic models. a) Model 1: A three-parameter model of expansion featuring
682 colonization of new population 2 by 2 diploid individuals from population 1 at time T_{EXP} . Population 1 is
683 of constant size N_1 , whereas population 2 grows exponentially to size N_2 , its size at present. b) Model 2:
684 the number of founders of population 2 is a variable parameter. c) Model 3: a per-generation migration
685 rate from population 1 to population 2 is added as a parameter. d) Model 4 includes all 5 parameters:
686 N_1 , N_2 , T_{EXP} , N_{O2} , and m_{21} .

687
688
689 **Figure 2.** Relative prediction error (RPE) calculated from the results of ABC analyses of 20 different
690 combinations of demographic models and sampling designs (x-axis). For each combination, ABC was
691 performed on simulated datasets summarized with statistics including linkage-based measures (hap.
692 phase 1) and on the same set of simulations summarized with only SNP-based statistics (hap. phase 0).
693 RPE values were calculated from the ABC estimation results of 1000 datasets with parameter values
694 randomly drawn from their prior distributions.

695
696
697 **Figure 3.** Width of the 95% highest posterior density intervals calculated from the results of ABC
698 analyses of 20 different combinations of demographic models and sampling designs. Error bars
699 represent standard errors (N=100 PODs). See caption of figure 2 for more details.

700
701
702 **Figure 4.** Accuracy of parameter estimates for model 1. Within a plot, each datapoint corresponds to
703 the estimated value of the parameter (mode of the posterior) vs. the true parameter value for one POD.
704 Results are shown for a total of 100 PODs. Error bars correspond to the 95% HDI around the estimate. a)
705 Results with datasets of type 1 (10,000 sequences of 100bp). Top panel shows results on unphased
706 datasets, bottom panel shows results for phased datasets. b) Results with datasets of type 5 (100
707 sequences of 10,000bp). Top panel shows results on unphased datasets, bottom panel shows results for
708 phased datasets.

709
710
711 **Figure 5.** RPE of model parameters for different fixed values of T_{EXP} . Results are shown for ABC runs with
712 datasets of type 1 (10k sequences, 100-bp long). For a given parameter, results from different models
713 are shown in the same plot window with different characters and colours. To see results for other
714 model-dataset combinations as well as 95% HDI results, please see supporting information.

715
716 **Figure 6.** RPE and bootstrapped confidence intervals of model 2 parameters under different sequencing
717 strategies and per-nucleotide error rates. N corresponds to the number of diploid individuals sequenced,
718 dp to the mean individual sequencing depth. “perf” corresponds to perfect datasets whereas “err0”,
719 “err0.001” and “err0.01” correspond to datasets where the sequencing process was simulated, with
720 depth sampling and errors introduced at rates 0, 0.001, and 0.01 substitutions per nucleotide
721 respectively. 13 PODs were used for each treatment.

722

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

726
 727
 728 **Figure 8.** Width of the 95% HDI from ABC results, compared to 95% CI from the SFS inference method.
 729 For each of the four demographic models, the same 10 simulated datasets were used as pseudo-
 730 observed datasets for both the ABC and the SFS runs. HDI and CI widths were calculated from 100
 731 bootstraps. Numbers correspond to the coverage of 95% CI (out of 10 PODs). PODs had 10,000
 732 sequences of 100bp genotyped in 20 diploid individuals.

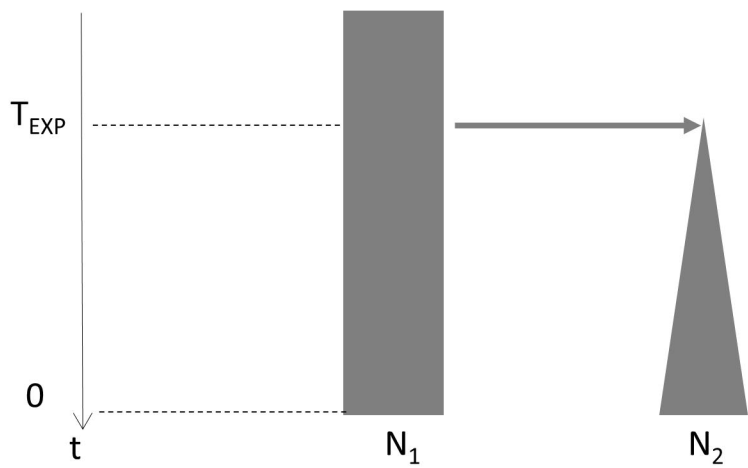
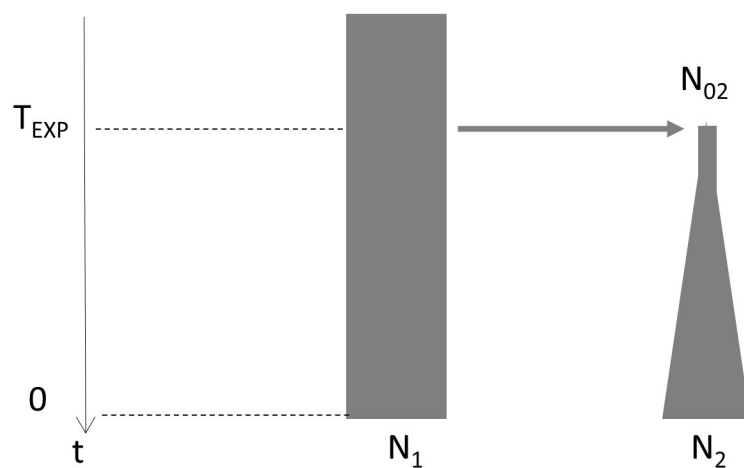
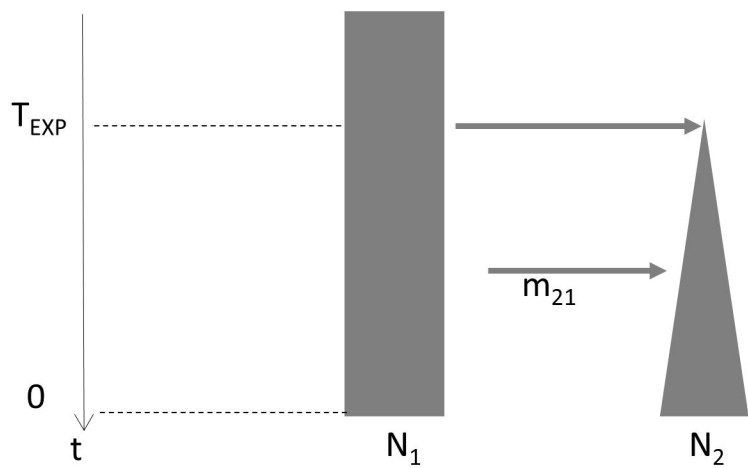
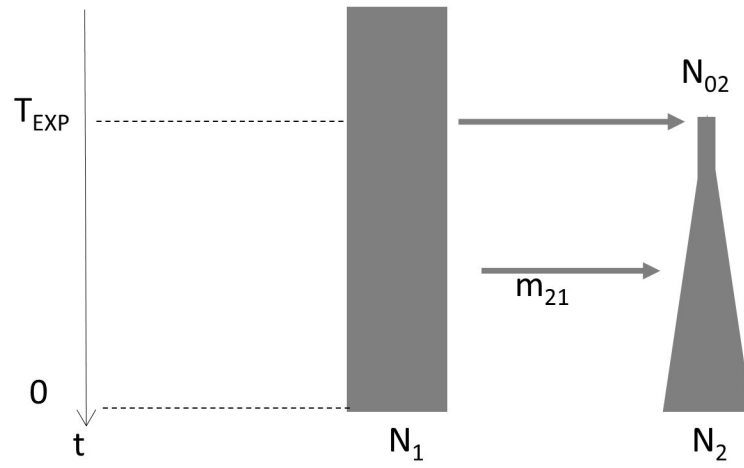
733
 734
 735 **Table 1.** Model parameters with their associated prior ranges

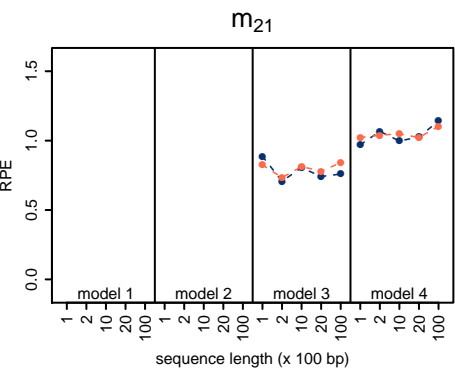
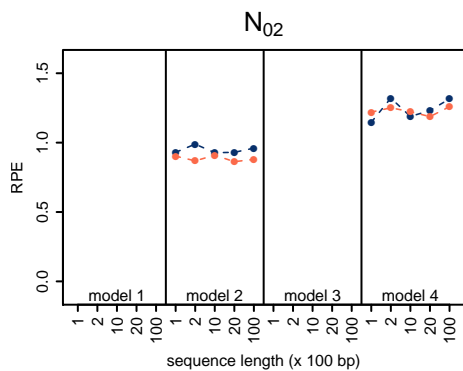
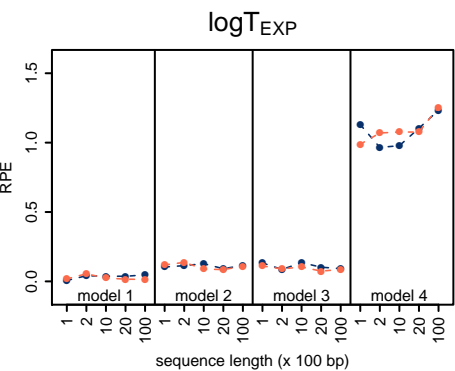
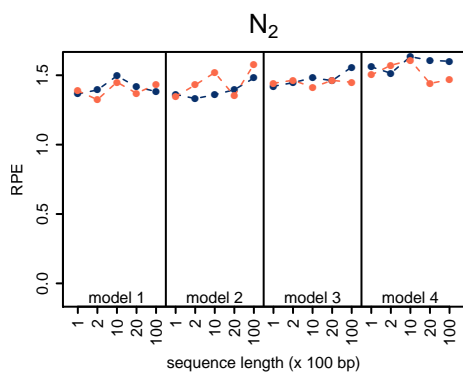
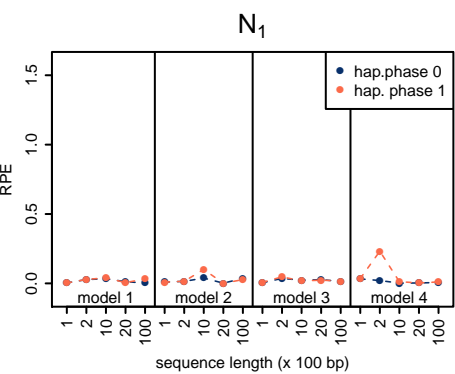
estimated in models	Parameter	Symbol	Prior range	Unit
-	Mutation rate	μ	0.000000009	-
-	recombination rate	R	0.00000001	-
1,2,3,4	population size 1	N_1	U(10,000:100,000)	ind.
1,2,3,4	population size 2	N_2	U(10,000:100,000)	ind.
1,2,3,4	time of expansion	T_{EXP}	logU(2:10,000)	gen
2,4	initial population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

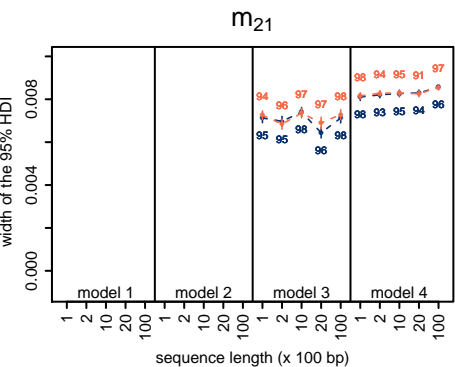
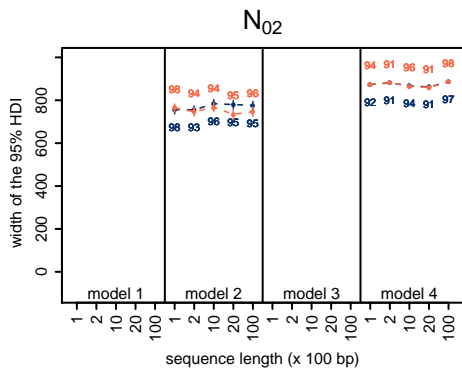
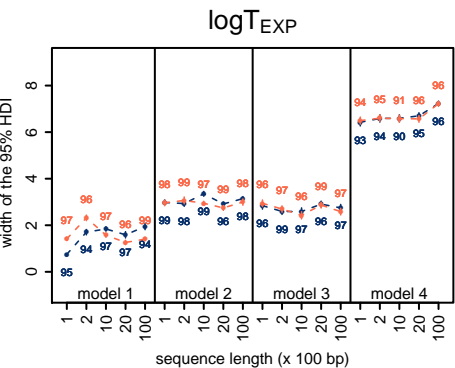
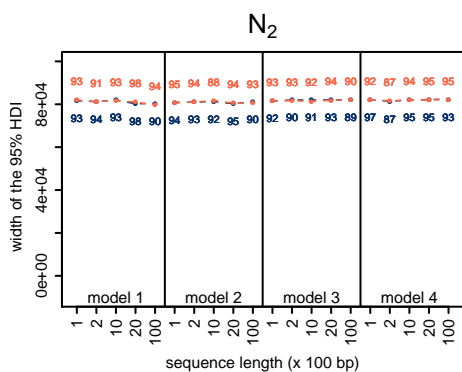
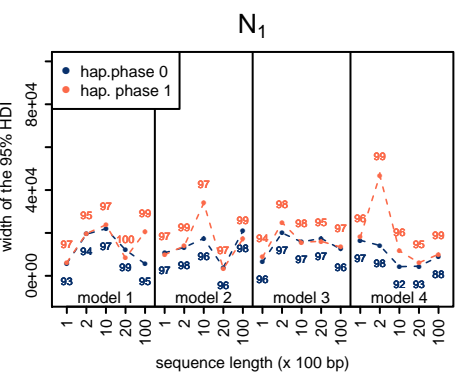
736
 737
 738
 739 **Table 2.** Description of the 5 types of simulated datasets

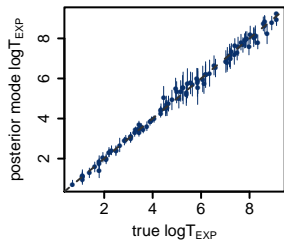
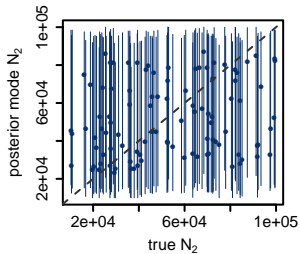
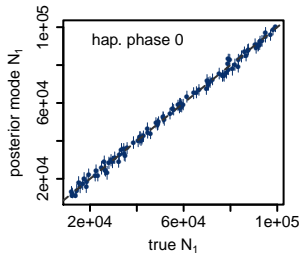
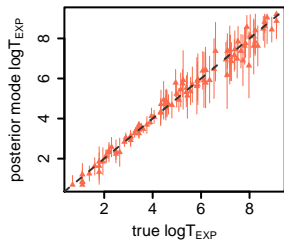
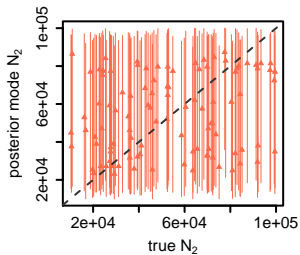
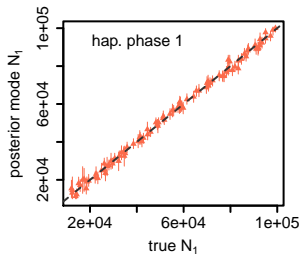
	number of sequences	sequence length (bp)	number of diploid individuals
1	10,000	100	20
2	5,000	200	20
3	1,000	1,000	20
4	500	2,000	20
5	100	10,000	20

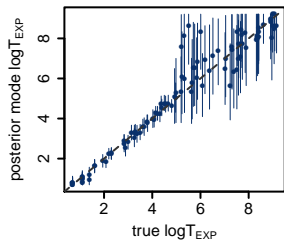
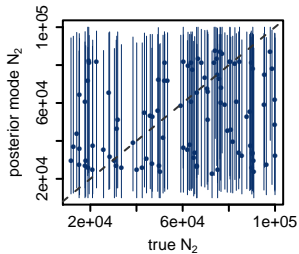
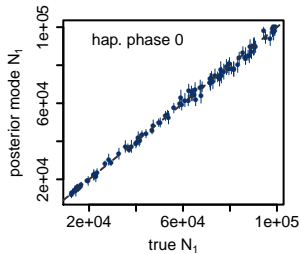
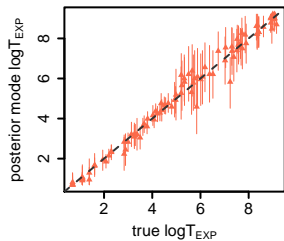
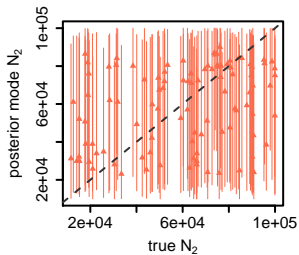
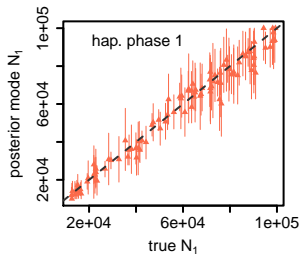
740

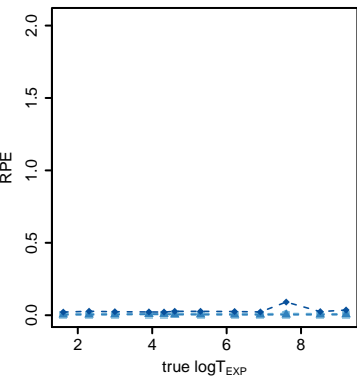
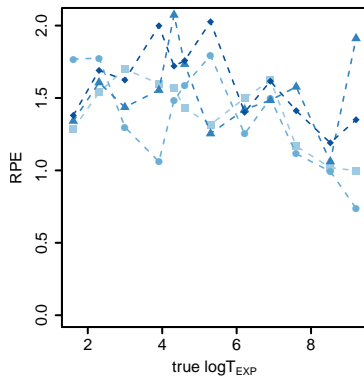
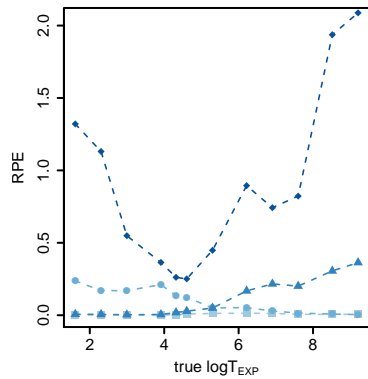
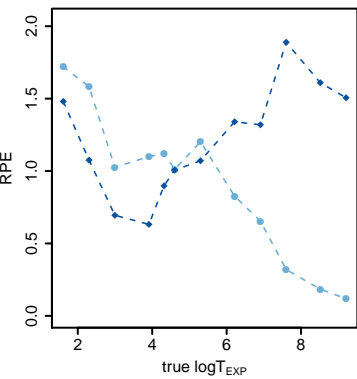
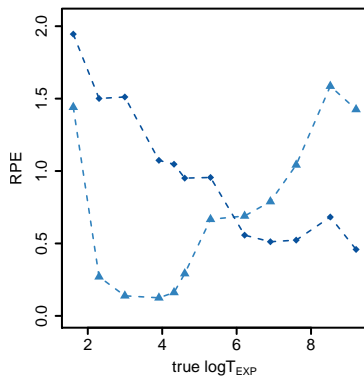
a)**b)****c)****d)**



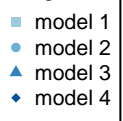


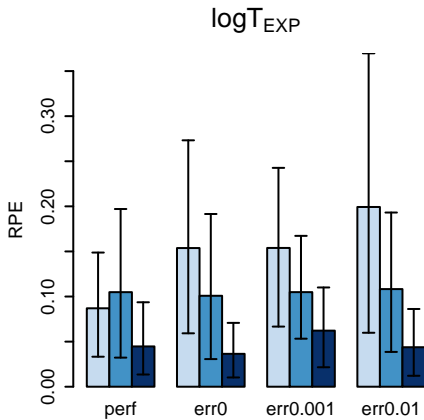
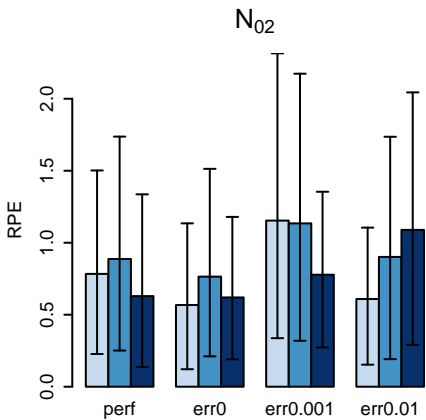
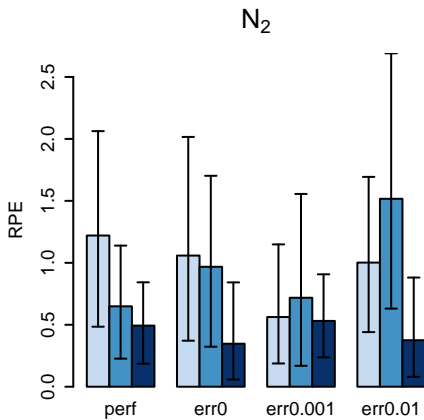
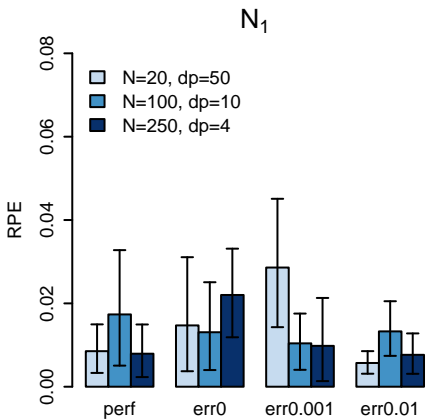
a.

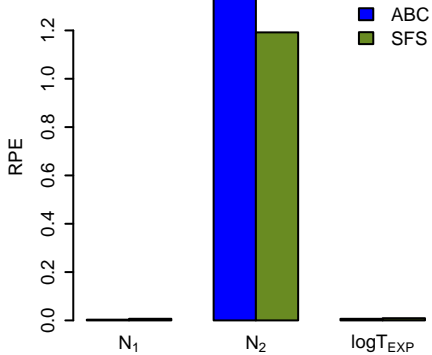
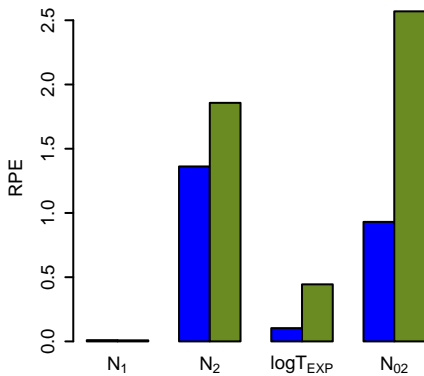
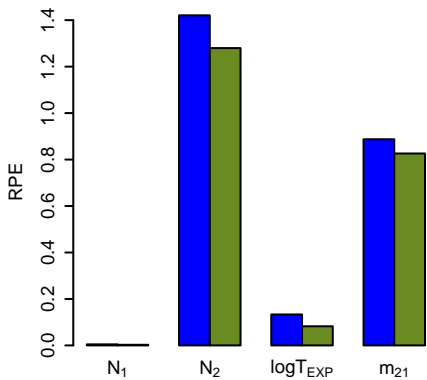
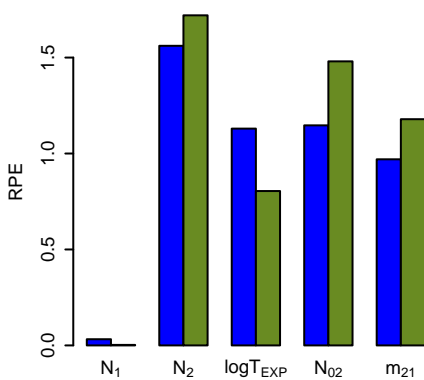
b.

N_1  N_2  $\log T_{EXP}$  N_{O_2}  m_{21} 

legend





model 1**model 2****model 3****model 4**

model 1

model 2

model 3

model 4

