Exploring Approximate Bayesian Computation for the inference of recent demographic history in non–model species with genomic markers

Journal:	*Molecular Ecology Resources*
Manuscript ID	MER-17-0254
Manuscript Type:	Resource Article
Date Submitted by the Author:	08-Aug-2017
Complete List of Authors:	Elleouet, Joane; University of British Columbia, Forest and Conservation Sciences Aitken, Sally; University of British Columbia, Forest and Conservation Sciences
Keywords:	approximate Bayesian computation, demographic inference, coalescent simulations, spatial expansion, population genetics

1   **Title**

2   Exploring Approximate Bayesian Computation for the inference of recent demographic history in non-

3   model species with genomic markers

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

Abstract

21

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 a class

24   of sequencing methods called reduced representation library (RRL), which selects a fraction of the

25   genome using targetted sequence capture or restriction enzymes (Genotype-by-sequencing, GBS). The

26   influence of the amount of information present in RRL datasets on the quality of demographic inference

27   needs assessing. Here, we explored the influence of marker number and length and knowledge of the

28   gametic phase on the quality of demographic inference performed with ABC. We focused on 2-

29   population models of recent spatial expansion with varying number of unknown parameters.

30   Performing ABC on simulated datasets with known parameter values, we found that the timing of a

31   recent spatial expansion event can be precisely estimated in a 3-parameter model. Taking into account

32   uncertainty in additional parameters such as initial population size and migration rate decreased the

33   precision of the inference dramatically for all parameters except current population sizes. Phasing

34   haplotypes did not improve results, regardless of sequence length, and numerous short sequences were

35   as valuable as fewer, longer sequences. ABC results were similar to results obtained with an alternative

36   method based on the site frequency spectrum (SFS) when performed with unphased GBS-type markers.

37   We conclude that unphased GBS-type datasets can be sufficient to precisely infer simple demographic

38   models, and discuss possible improvement for the use of phased haplotypes.

39

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   using sophisticated techniques for marker development. In parallel, inferring complex 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 different levels of completeness of genomic datasets

53   need to be understood to ensure quality and accuracy of inferences accompany this trend.

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 manage

64   invasive species (Benazzo et al., 2015; Guillemaud et al., 2010), to conserve endangered species or

65   ecosystems (Chan et al., 2014; Dussex et al., 2014; Lopez et al., 2006; Quéméré et al., 2012), and to

66   predict the future distribution and abundance of widespread species that are of economical or ecological

67 importance (Holliday et al., 2010; Zinck & Rajora, 2016). The good news is the genomic revolution has

68 reached a number of non-model organisms, creating a spectrum of levels of genetic knowledge across a

69 broad range of taxa. Using a few microsatellites or moderate-sized panels of resequenced SNPs is still

70 common practice (Li et al., 2010; Zinck & Rajora, 2016), but most current studies of non-model species

71 now use genomic methods to extract markers for inference. Sequencing whole genomes of non-model

72 species has become feasible in some organisms with small genomes in recent years (Boitard et al., 2016;

73 Liu et al., 2014) and has allowed the inference of detailed demographic models using ABC or the Pairwise

74 Sequential Markovian Coalescent (PSMC) technique (Nadachowska-Brzyska et al., 2013). For organisms

75 with larger genomes or for studies with lower requirements for data quantity, reduced-representation

76 library sequencing through targetted capture or restriction enzymes is widely applied. (Davey et al.,

77 2011). The set of techniques involving restriction enzymes (commonly referred to as RADseq or

78 Genotyping-by-sequencing, GBS) outputs a large number of short sequences (100bp, or longer with

79 paired-end sequencing) from across the genome and has proven useful in population genetics studies

80 and inference involving maximum likelihood methods based on the site frequency spectrum (SFS) or ABC

81 methods (Narum et al., 2013). Most recently, the number of published drafts of whole genomes for non-

82 model species has increased dramatically, granting access to longer sequences through the second

83 category of genomic markers: targeted enrichment. This approach allows the use of linkage information

84 for population genetics inference (Li & Jakobsson, 2012).

85

86 **Approximate Bayesian computation (ABC) and other approaches**

87   In this paper, our aim is to explore the ABC method for datasets obtained from reduced-

88 representation library sequencing in non-model organisms. We also compare the obtained results with

89 results from a SFS approach based on the approximation of the composite likelihood (Excoffier & Foll,

90   2011). We chose to explore ABC because its versatility accomodates a wide spectrum of demographic

91   models and dataset types. Although it was originally developed for inferences in evolutionary biology,

92   the statistical framework of ABC has been extended to a variety of disciplines, from cell biochemistry and

93   epidemiology to neural networks, extending beyond the realm of biology into meteorology, astrophysics

94   (Weyant *et al.* 2013) and computer sciences (Condon & Cukier, 2016). The ABC method has been

95   reviewed in a number of publications and its algorithms and techniques are being refined constantly

96   (Bertorelle et al., 2010; Csilléry et al., 2010; Lintusaari et al., 2016; Marin et al., 2012; Sunnaker et al.,

97   2013). For applications in demographic inference using genetic data, the general ABC method involves

98   the following steps. First, A large number of datasets are simulated under a specific demographic model

99   using the coalescent (Kingman, 1982). Parameters used for simulations are drawn from prior

100   distributions that are pre-defined by the user. The simulated datasets are then compared to the

101   observed dataset through calculation of carefully chosen summary statistics. Finally, simulated datasets

102   with the closest vector of statistics to the vector of observed summary statistics are selected and their

103   original parameter values are used to approximate the posterior distribution of each model parameter.

104   The ABC method is suitable when inferring models for which the likelihood function is intractable, as it

105   relies on approximating the likelihood function using a large number of simulations. However, each one

106   of the numerous steps in the implementation of ABC required empirical decisions to be taken by the

107   user. Especially, there is a need to improve our understanding of the relationship between the type of

108   markers obtained to build genetic datasets, the way genetic data is subsequently summarized, and its

109   power to tease apart demographic models and produce accurate parameter estimates.

110

111   **Previous work exploring the ABC method**

112   The need to test the inference power of datasets for demographic models of interest has been

113   recognized in recent years. Several studies show the use of preliminary simulations testing the number

114   and length of markers and the number of individuals (Sousa et al., 2012; Stocks et al., 2014), the type of

115   molecular markers (Cabrera & Palsbøll, 2017) and the choice of summary statistics and models

116   considered (Benazzo et al., 2015; Guillemaud et al., 2010; S. Li & Jakobsson, 2012; Sousa et al., 2012;

117   Stocks et al., 2014). As most scientists have switched to using genome-wide data, there is a need to

118   expand this set of simulation studies to test and understand the power of different types of genomic

119   data. As part of such an effort, Li & Jakobsson (2012) simulated large, phased genomic datasets

120   comparable to human genomic datasets at the time. Under 2-population split models, they found that

121   ABC produces accurate estimates for most but not all parameters and concluded ABC is well suited to

122   large genomic datasets summarized with LD-based statistics. Robinson et al. (2014) tested the effects of

123   the number and length of unphased genomic sequences and compared them to the effect of the

124   number of individuals sequenced for the inference of three-population admixture models. They found

125   that increasing the number and length of sequences was more beneficial than increasing sample size.

126   Shafer et al. (2015) investigated the power of ABC on short diploid sequences obtained by GBS. They

127   focused on a wide range of simple 1-population and 2-population models with bottleneck, growth,

128   migration and a combination of these parameters. They found that population changes such as ancient

129   temporary bottleneck would not be inferred correctly regardless of the number of markers available.

130   This set of studies provides valuable information about the use of genomic data in ABC. Our aim is to

131   extend this knowledge by directly comparing ABC results from molecular markers obtained with

132   different types of RRL sequencing techniques and with different levels of genomic knowledge. This will

133   hopefully help future ABC users who do not have access to complete genomic data to select methods

134   and develop genomic datasets that are best suited to answer the demographic questions they are

135   addressing.

136

General model and datasets

138   Here, we focus on a set of 2-population models of demic expansion that are applicable to studies

139   of species invasion, reintroduction, or natural colonization. We tested the power of ABC on these models

140   using a range of marker sets obtainable by RRL methods: datasets with a large number of short genomic

141   reads would correspond to single-end GBS sequencing, whereas fewer but longer diploid sequences

142   correspond to a targeted enrichment approach. For each type of dataset, we quantified the potential

143   benefits of knowing the gametic phase of sequence markers by including or excluding linkage-related

144   statistics at the data-summarizing step. We expected to observe an improvement in the inference for

145   datasets with long sequences. For each model assessed, we also tested the effect of time since the

146   colonization event occurred. We hypothesize that recent events might be inferred more accurately with

147   datasets containing linkage information, due to the generally higher rate of recombination compared to

148   mutation, and to the potential information contained in long haplotypes. This part of the analysis is also

149   motivated by the fact that overestimates of divergence times are a common result of demographic

150   inference in empirical studies (Holliday et al., 2010) and this upward bias has been found for some

151   demographic scenarios in simulation studies (Benazzo et al., 2015). We therefore aim to explore this

152   potential bias by testing increasingly old events within the same models. Finally, we compared our ABC

153   results with those obtained from an approximate likelihood method using the site frequency spectrum

154   from simulated reduced-representation libraries. As they provide millions of genome-wide SNPs without

155   ascertainment bias, restriction enzyme-based genomic sequencing techniques seem to be particularly

156   well suited to SFS-based inference methods. Comparing SFS results with ABC results on a range of

157   models and datasets will inform future work on demographic inference in non-model organisms.

158

159   **Methods**

160   **Demographic models**

161   We focused on a basic 2-population model of demic expansion (fig.1a). A pre-existing

162   population, population 1, is of constant size N_1. At time T_{EXP} before present, the spatial population

163   expansion begins: population 2 is created by 2 migrants from population 1. Population 2 then grows

164   exponentially between times T_{EXP} and T=0 (present) to size N_2. The rate of expansion is defined by the

165   other parameter values through the formula $r=\log(\frac{N_{02}}{N_2})/T_{EXP}$. Model 1 therefore has just 3 independent

166   unknown parameters: N_1, N_2, and T_{EXP}. We created additional models of increasing complexity by adding

167   parameters. In models 2 and 4, the number of founders of population 2, N_{02}, is unknown (fig.1b and

168   fig.1d); in models 3 and 4, migration is allowed from population 1 to population 2, with the parameter

169   m_{21} describing a per-generation migration rate (fig.1c and fig.1d). In all four models described above,

170   mutation and recombination rate are fixed. Parameter priors were chosen to approximate studies of

171   postglacial expansion of temperate tree species (Table 1); population sizes are generally large and

172   expansion time has an upper limit of 500 generations. However, we set priors with wide ranges so that

173   results would inform students of other biological systems.

174

175   **Genomic datasets simulated**

176   For each of the four models, we created a set of 1 million simulations with each of the five types of

177   datasets described below, with a fixed number of 10 diploid individuals sampled per population. For

178   datasets corresponding to single-end RADseq sequencing techniques, we simulated 10,000 independent

179   DNA sequences of 100bp each. For datasets corresponding to sequence capture methods, we created

180   100 independent DNA sequences of 10kb each. Additionally, we explored a range of possible

181   configurations between these two types of datasets (table 2). With 4 models and 5 types of datasets, we

182   obtained a total of 20 combinations of models and datasets, each with a million simulations. We used

183   the program scrm (Staab et al., 2015) in R (R Core Team, 2016) to compute the simulations, using custom

184   scripts inspired by scripts from Shafer et al. (2015) and available in the supporting information.

185

186   **Summary statistics**

187   For each simulated dataset we used the program msABC (Pavlidis et al., 2010) to compute all

188   summary statistics available, including diversity statistics (number of segregating sites and θ estimates),

189   summaries of the SFS (Tajima's D and Fay and Wu's H), differentiation measures (pairwise F_{ST}, number of

190   private and shared polymorphisms), the Thomson estimator of T_{MRCA}, and its variance. To test the effect

191   that knowing haplotype information has on inference, the ABC analysis was performed twice on each

192   model-dataset type combination. The first time, we summarized data using only the statistics mentioned

193   above, which are calculated at the SNP level and therefore are available when the gametic phase of the

194   diploid sequences is unknown. The second inference was performed on the same dataset, but additional

195   statistics based on linkage information were used to summarize the data: Zns (Kelly, 1997), dvk and dvh

196   (Depaulis & Veuille, 1998). These additional statistics are calculated at the haplotype level, they are

197   therefore only available in cases where the gametic phase of the diploid sequences is known. For each

198   set of simulations, we computed the mean and variance of every statistic over all sequence markers in

199   the dataset. As a result, 55 statistics were computed for datasets with known gametic phases (hereafter

200   referred to as "phased", or "hap.phase 1"), and 40 statistics were computed for datasets with unknown

201   gametic phases (hereafter referred to as "unphased", or "hap.phase 0").

202   Using a high number of statistics to summarize genetic data has harmful effects on the quality of

203   the ABC inference, a problem commonly referred to as the "curse of dimensionality" (Blum et al., 2013).

204   We used the partial least squares (PLS) method implemented in ABCtoolbox (Wegmann et al., 2010) to

205   reduce the number of statistics to 5 PLS components (see Supplemental methods for details).

206

207   **ABC estimation**

208   We created independent pseudo-observed datasets (PODs) for each combination of dataset and

209   demographic model, using the exact same genetic settings and prior ranges of parameters as for the

210   corresponding set of 1 million simulations. We then performed the ABC estimation using each POD as

211   the observed dataset to evaluate precision and accuracy of estimates. The standard ESTIMATE algorithm

212   from the program ABCtoolbox (Wegmann et al., 2010) was used for all ABC computations, with a post-

213   sampling regression adjustment through ABC-GLM (Leuenberger & Wegmann, 2010). We fixed the

214   tolerance parameter to 10^{-3} , a compromise between having a tolerance threshold value as low as

215   possible (S. Li & Jakobsson, 2012) and keeping an appropriate number of simulations to estimate the

216   posterior from.

217

218   **Validation**

219   For each combination of model and type of dataset, we first performed a random validation step as part

220   of the estimation process. i=1000 datasets were simulated with random parameter values drawn from

221   the prior ranges, and parameters were estimated for each of them with unchanged settings. For each

222   parameter, we calculated the root mean squared error (RMSE) from the results of these i=1000 random

223   simulations, following equation (1):

224   (1) $\varepsilon = \sqrt{\dfrac{\sum_1^i(\hat{\theta}-\theta^*)^2}{i}}$

225   where θ^* is the parameter value used for the simulation and $\hat{\theta}$ is the estimate defined as the parameter

226   value with highest posterior probability.

227   For comparison purposes, we computed a complementary measure of precision and accuracy

228   called the relative prediction error (RPE), the ratio of the variance of the posterior over the variance of

229   the prior, which follows equation (2):

230   (2) $\varepsilon = \dfrac{\sum_1^i(\hat{\theta}-\theta^*)^2}{Var(\theta^*)} \times \dfrac{1}{i}$

231   where $Var(\theta^*)$ is the variance of the prior distribution. The advantage of using RPE as a validation

232   statistic is that it directly indicates the contribution of the genetic dataset to the estimation of the

233   posterior. Another attractive feature of the RPE is that it allows comparisons between parameters, as it

234   scales from 0 (precise estimate) to 1 and beyond (in the case of a consistent bias in estimation).

235   The 95% highest posterior density interval (HDI) was calculated as an additional measure of

236   precision. It is defined as the shortest continuous interval with an integrated posterior density of a

237   certain value (Wegmann et al., 2010). As such, it is an ABC equivalent of the Bayesian credible interval.

238   The 95% HDI was calculated on 100 PODs for each set of 1M simulations.

239

240   **Testing the effect of T$_{EXP}$ on the parameter estimation**

241   To test the effect of the time of expansion on the precision of the ABC estimation, we created

242   100 PODs for each set of 1M simulations and each of the following T$_{EXP}$ values: 2, 5, 10, 20, 50, 75, 100,

243   150, 200, 300, 400 and 500 generations. RMSE and 95%HDI were calculated from the results of each set

244   of 100 PODs.

245

246   **Comparing ABC and SFS estimation**

247   We simulated 10,000 independent DNA sequences of 100bp each for the 4 demographic models 10

248   times. The resulting 40 datasets were input into both ABCtoolbox and fastsimcoal2, which uses the SFS

249   to approximate a composite likelihood from a large number of simulations through a conditional

250   maximization algorithm (see supplemental materials and methods). We compared the results from the

251   two methods using RPE, RMSE and confidence intervals.

252

253   # Results

254   A total of 20 combinations of models and datasets were used as input for ABC simulations

255   (tables 1 and 2), resulting in a total of 20 million simulated datasets available for analysis, training

256   simulation sets and PODs. Each set of 1M simulations was used in two runs of estimation: one including

257   all summary statistics available in msABC, the other one excluding statistics based on linkage

258   information, for a total of 40 ABC estimations.

259

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

261   In general, the ability to infer demographic history declined rapidly as model complexity

262   increased. The simplest model (model 1), estimating only both population sizes N_1 and N_2, and the time

263   of expansion T_{EXP}, allowed dating the expansion event accurately. Models 2 and 3 each had 4

264   parameters: model 2 included the number of founders N_{02} and model 3 allowed migration from

265   population 1 to population 2 (m_{21}). Both models were inferred with moderate precision. Finally,

266   scenarios corresponding to model 4, which had all 5 parameters, failed to be correctly inferred.

267   Not all parameters were sensitive to the addition of parameters in the models: the precision of

268   contemporary population size estimates N_1 and N_2 was independent of model complexity. RMSE values

269   for N_1, which were set as constant over generations, were below 10,000 for the four models assessed

270   (fig.2). 95% highest posterior density intervals ranged from 1000 to 4000. For N_2, the contemporary

271   population 2 size after exponential growth, 95% HDI intervals were about as wide as the prior range,

272   indicating a failure to estimate this parameter in all four models.

273   The expansion time T_{EXP} was generally well estimated in model 1, which is the simplest, 3-

274   parameter model (fig.2) with no migration between demes and the number of founders set to 2. For this

275   model, 95% HDI intervals show moderate precision with ranges as wide as 150 to 200 generations. The

276   precision of the T_{EXP} estimation was mediocre for more models where the number of founders N_{02}

277   (model 2) is unknown and needs to be estimated, or where migration from population 1 to population 2

278   is likely (model 3). For these two models, RMSE values of T_{EXP} were around 110, and 95% HDI were as

279   wide as 300 generations. The ABC analysis on the 5-parameter model (model 4) was unable to recover

280   the true T_{EXP} value (95% HDI ≥ 400).

281   Estimates of the number of founders of population 2 (N_{02}) and migration rate from population 1

282   to 2 (m_{21}) were also impacted by model complexity. In 4-parameter models, RMSE values for N_{02} were

283   between 220 and 240, and RMSE values for m_{21} were between $1.7\ e^{-3}$ and $2.4\ e^{-3}$. N_{02} values could not be

284   recovered in the most complex model 4, and m_{21} RMSE values were high. Generally, the 95% highest

285   posterior density intervals followed the same trends as the RMSE for N_{02} and m_{21} (fig.3): N_{02} and m_{21}

286   were each impossible to infer in the presence of the other.

287   Our models all rely on population 2 growing exponentially from T_{EXP} to the present time. We tested

288   whether demographic parameters could be estimated more successfully in a model where population 2

289   goes through a single sudden population change instead of an exponential growth. We created a new set

290   of 1M simulations based on model 2 (where N_{02} is a varying parameter) and dataset type 1 (many short

291   sequences). In the new model the size of population 2 changes from N_{02} to N_2 at $T_{EXP}/10$ and remains

292   constant before and after $T_{EXP}/10$. The modification brought no improvements to any of the parameter

293   estimates (Table S1).

294

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

296   The addition of the linkage statistics available in msABC brought no notable improvement in the

297   RMSE and 95% HDI of parameter estimates for all models (fig.2 and fig.3). It even seems to make the

298   estimation of N_1 less precise in model 1, as shown by an increase of up to 36000 of 95% HDI values

299   (fig.3). ABC performance on models 3 and 4 seemed to be slightly more dependent on sequence length,

300   with the inference on large sequences marginally benefitting from haplotype information.

301

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

303   For each parameter, we visualized estimated values and 95% HDI of ABC results in relation to

304   true parameter values to assess performance over the prior range. Results for the 3-parameter model

305   (model 1) and dataset types 1 and 5 are shown in fig.4a and fig.4b, respectively. Results for the complete

306   set of models are available in supplemental fig.S1. Consistently across models, estimates of N_2 are

307   largely inaccurate regardless of the true value, with HDI ranges as wide as the prior range. Conversely, N_1

308   estimates are accurate in all models regardless of the true N_1 value. Unlike N_1 and N_2, the values of T_{EXP},

309   N_{02} and m_{02} do have an impact on the precision of their respective estimates. Accuracy and precision of

310   T_{EXP} estimates for models 1 to 3 decrease with increasing true value. N_{02}, the number of founders of

311   population 2, is estimated in model 2 (4 parameters) and model 4 (5 parameters). Inferences on model 4

312   failed to estimate N_{02} across all dataset types and haplophases. For inference of model 2, estimates of

313   N_{02} values below 150 are fairly accurate, with 95% HDI intervals as low as 20 when simulating datasets

314   with marker lengths 100 and 200. With datasets of fewer, longer sequences, a smaller range of low N_{02}

315   true values lead to a high level of precision. Finally, the per-generation migration rate from population 1

316   to 2, $m_{21,}$ is simulated in models 3 and 4. Its estimation follows a similar trend as T_{EXP}: the larger true m_{21}

317   values are, the less accurate and precise their estimates become (fig.S1). m_{21} values below 0.002 could

318   be estimated with moderate precision in model 3, and very imprecisely in model 4 (fig.S1).

319

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

321   We tested whether older expansion events are generally more difficult to characterize than

322   recent ones within the time range specified by the prior. To do this, we studied the effect of the true T_{EXP}

323   value on the precision of parameter estimates. We find different trends foramong the 4 models (fig.5,

324   S2, and S3). The precision of inference on model 1 is high at low T_{EXP} values and decreases as T_{EXP}

15

325   increases. For model 2, older events seem to be generally better inferred. However, the precision of T_{EXP}

326   estimantes decreases with T_{EXP} but the precision of N_{02} has the opposite trend, suggesting that

327   estimating both T_{EXP} and N_{02} precisely is almost impossible. For both model 1 and 2, there is a decrease

328   in RMSE and 95% HDI estimates at high T_{EXP} values, but this probably only reflects the fact that

329   prediction values are bounded to the upper prior limit (fig.S2 and S3). Model 3 shows best results for

330   moderately recent expansion events (20-75 generations), as shown by estimates of T_{EXP} and m_{21}. Finally,

331   results for model 4 show high values of RMSE and 95% HDI for all parameters, but it seems that the

332   precision of T_{EXP}, N_{02}, and m_{21} is generally higher for older expansion events.

333   Looking at results for individual parameters across all models, we note that the estimation of T_{EXP}

334   is better for a more recent expansion event as long as there is no migration between populations (fig.5).

335   The precision of N_{02} estimates is highly dependent on T_{EXP}, especially for model 2, where RMSE values

336   decrease from 400 to as low as 100 as T_{EXP} increases. Estimation of the migration parameter m_{21} also

337   seem to be affected by the timing of population expansion but the relationship is complex in model 3. In

338   model 4, RMSE values for m_{21} decrease as T_{EXP} increases.

339

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

341   Figures 6 and 7 illustrate the performance of ABC and approximate composite likelihood from

342   the SFS for all models performed with datasets of 10,000 100-kb sequences. Both methods gave similar

343   results in terms of precision of parameter estimates. The SFS-based method performed slightly better

344   than ABC in the model with migration (model 3), although this superiority is not observed in the width of

345   confidence intervals. ABC resulted in a lower relative prediction error for model 2 in the estimation of

346   N_{02} (fig.6), but again the width of confidence intervals was similar across methods.

347

348

Discussion

350   We explored the ability of approximate Bayesian computation to characterize a recent event of

351   spatial expansion from one population of constant size to a new growing population, a model which can

352   be broadly applied to studies of species range expansion, invasion biology, or reintroduction of

353   endangered species. Our results show that higher model complexity provided results of lower quality.

354   Precisely dating a spatial population expansion event that occurred in the recent past was not possible

355   for complex models (more than 4 parameters). The simplest model with only three parameters (N_1, N_2

356   and T_{EXP}) could be partially recovered with a precise estimate of timing of expansion and size of source

357   population (N_1, assuming it is constant) but a poor estimate of the size of the growing, newly founded

358   population (N_2). Failure to estimate N_2 does not come as a surprise: estimates of past changes in

359   effective population size from one punctual sampling event commonly rely on linkage information

360   between markers, a calculation not readily available in ABC packages (Beaumont, 2003; Waples & Do,

361   2008). The inference of models including the number of founders of population 2 (N_{02}) and/or per-

362   generation migration rate from population 1 to population 2 (m_{21}) was only partially successful. The

363   number of founders and time of founding could not be jointly estimated precisely. We found that the

364   timing of a demographic event has an effect on the precision of parameter estimates with ABC: the time

365   of expansion itself was more precisely estimated when it was moderately recent, a trend also observed

366   in Li and Jakobsson (2012). The presence of migration makes inference of the time of expansion

367   impossible when it is more recent than about 10 generations, and historical events could only be

368   inferred for a narrow time range (20 to 50 generations ago). The difficulty of estimating the time of a

369   population founding event with subsequent migration has also been reported in Robinson et al. (2014).

370   The number of founders of population 2 and the per-generation migration rate from population 1 to

371   population 2 were both estimated more precisely for events that are more ancient.

372

373   **Implications of including haplotype information**

374   Analyses based on unphased sequences exploring similar models as here has shown encouraging

375   results (Robinson et al., 2014). However, no study to date has explicitly compared datasets of phased

376   and unphased sequences using the same models and same amounts of data. Here, we quantified the

377   benefits of using haplotypes of phased sequences over single SNPs by including or leaving out LD-based

378   and haplotype-level statistics at the data summarization step of the ABC inference. Adding haplotype

379   information to simulated datasets made surprisingly little difference compared to identical ABC runs

380   with phased data. The reason why phasing the data did not improve inference could be that the extent

381   of linkage the chosen statistics are sensitive to is different from the linkage actually present in the

382   simulated data. In the present study, we decided to use all statistics available in msABC. We chose to

383   simulate sequences with a fixed per-nucleotide recombination rate r of 10^{-8}, a value compatible with

384   estimates in mammal species but much lower than estimates in other organisms such as plants. We

385   simulated sequences spanning a maximum of 10^4 bp. This results in a maximum per-marker

386   recombination rate of 10^{-4} per generation. Considering that the effective population sizes simulated

387   were between 10^4 and 10^5, the population-wise per-generation recombination rate for a 10kb marker

388     $4N_e r$ could be as high as 40 per generation. For a demographic event of at least a few generations,

389   recombination rates of this order of magnitude should at least affect the variance of LD-based summary

390   statistics, although this depends on the model considered. Future work when dealing with phased data

391 requires developing expectations of LD levels and creating or choosing statistics that cover the extent of

392 LD likely to be present in the data. ABC on phased data also requires decent knowledge of the

393 recombination rate and its variability across the genome so that this information can be included in the

394 simulations. The recombination rate therefore needs to be included as a parameter along demographic

395 parameters, or as a nuisance parameter with a hyper-prior. As for mutation rates, different ways to deal

396 with them in ABC analyses have been explored in Shafer et al. (2015).

397 Li and Jakobsson (2012) explored the ABC method with similar, 2-population split models and a

398 similar, fixed population-wise per-generation recombination rate as our study. When they tested

399 different combinations of summary statistics, their results did not demonstrate any obvious superiority

400 of LD-based statistics over SNP-based statistics. They concluded that chosen summary statistics should

401 capture as many different aspects of the data as possible, with as little redundancy as possible.

402 On a practical note, one needs to be aware that simulating the coalescent with recombination is

403 a complicated process and comes at high computational costs (McVean & Cardin, 2005). With high

404 recombination rates and/or very long sequences, coalescent simulations might take so long to run that

405 one would likely chose a more efficient inference method than ABC. Moreover, translating genome-wide

406 observed data into a set of summary statistics values readily useable by ABC programs and comparable

407 to simulated datasets can be a challenge. For example, when aligning reads to a fragmented and

408 incomplete reference genome, as is often the case in non-model organisms, defining haplotypes can be

409 tricky. One needs to tackle the problems of sequencing errors, paralogous sequences and imperfect

410 mapping. Inevitable sequencing uncertainties will affect haplotype statistics more strongly than they

411 affect single-SNP diversity measures. Data processing errors and filters do severely bias the inference, to

412 the extent of supporting the wrong demographic model, as revealed by Shafer et al. (2016). Finally,

413 targeted sequence capture will result in a set of thousands of markers of various lengths. Defining a set

414   of millions of simulations that correspond closely to an observed dataset requires approximating the

415   distribution of sequence lengths, and this may also affect inferences, especially if the variance of

416   summary statistics are included at the data summarization step.

417

418   **Data structure and quantity**

419   We kept constant the total amount of sequenced data over all five types of datasets. Our aim

420   was to focus on other, less studied aspects of data collection design. Studies exploring the effects of the

421   amount of data in ABC inference have generally found that increasing the number of genetic markers is

422   more efficient than increasing the number of individuals sampled (Robinson et al., 2014; Stocks et al.,

423   2014). These studies found a threshold number of markers beyond which the precision of parameter

424   estimates and of model choice reach a plateau. Both Robinson et al (2014) and Li & Jakobsson (2012)

425   have concluded 2000 loci is an optimal number, although this would depend on many other variables

426   such as locus length, model complexity, and even true value of parameters (Shafer et al., 2015).

427

428   **Comparing ABC to other methods**

429   We did not find large differences in the precision of parameter estimates between ABC and the

430   SFS-based likelihood method implemented in fastsimcoal 2. Shafer et al. (2015) found a similar result

431   while comparing the performance of ABC with a SFS-based inference implemented in $\delta a \delta i$ (Gutenkunst

432   et al., 2009). They found that $\delta a \delta i$ tends to overestimate the time of population split and bottleneck

433   events, a trend not supported by our findings with *fastsimcoal*. In addition to parameter estimation,

434   Shafer *et al.* (2015) tested the performance of both methods for model selection and found ABC more

435   accurate, especially in the case of bottleneck scenarios.

436   ABC has proven moderately useful for demographic inference with long, genome-wide

437   haplotypes but comparisons with alternative approaches is scarce. Notable examples include

438   Nadachowska-Brzyska et al. (2013), who used ABC and PSMC in a complementary way; Robinson et al.

439   (2014) compared their ABC results with an exact likelihood method developed by Lohse et al. (2011) and

440   found that ABC resulted in more uncertainty, especially in model comparisons. As ABC performance with

441   linkage information needs to be further explored, comparisons to emerging analytical methods based on

442   whole genomes or long sequences such as MSMC (Schiffels & Durbin, 2014) or Identity-by-descent

443   haplotype sharing (Harris & Nielsen, 2013) will greatly help define the future of demographic inference

444   using data at a genomic scale.

445   Theoretical improvements of ABC methods continue to appear at a fast rate. For example, the

446   choice of summary statistics can be made objectively in increasingly fancy ways: Prangle et al. (2014)

447   recently proposed using a subset of the simulations to perform a regression of parameters on summary

448   statistics to define a projection matrix that is subsequently used to create linear vectors, and these in

449   turn can be used as summary statistics. One can also improve the estimation from simulations by having

450   a variable tolerance value that stretches along uninformative statistics and is reduced along informative

451   ones (Prangle, 2017). This versatility might be key to the success of ABC methods in a wide variety of

452   fields, even those experiencing rapid progress such as population genetics. We therefore predict that

453   ABC will continue to be used for demographic inference and adapt to the drastic change in the type of

454   genetic data becoming available.

455

460
461
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

504 Guillemaud, T., Beaumont, M. A., Ciosi, M., Cornuet, J.-M., & Estoup, A. (2010). Inferring introduction
505   routes of invasive species using approximate Bayesian computation on microsatellite data.
506   *Heredity*, *104*, 88–99. https://doi.org/10.1038/hdy.2009.92
507 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the Joint
508   Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS*
509   *Genetics*, *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695
510 Harris, K., & Nielsen, R. (2013). Inferring demographic history from a spectrum of shared haplotype
511   lengths. *PLoS Genetics*, *9*, e1003521–e1003521. https://doi.org/10.1371/journal.pgen.1003521
512 Holliday, J. a, Yuen, M., Ritland, K., & Aitken, S. N. (2010). Postglacial history of a widespread conifer
513   produces inverse clines in selective neutrality tests. *Molecular Ecology*, *19*, 3857–64.
514   https://doi.org/10.1111/j.1365-294X.2010.04767.x
515 Kelly, J. K. (1997). A test of neutrality based on interlocus associations. *Genetics*, *146*, 1197–1206.
516 Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, *13*, 235–248.
517   https://doi.org/10.1016/0304-4149(82)90011-4
518 Leuenberger, C., & Wegmann, D. (2010). Bayesian computation and model selection without likelihoods.
519   *Genetics*, *184*, 243–252. https://doi.org/10.1534/genetics.109.109058
520 Li, S., & Jakobsson, M. (2012). Estimating demographic parameters from large-scale population genomic
521   data using Approximate Bayesian Computation. *BMC Genetics*, *13*, 22–22.
522   https://doi.org/10.1186/1471-2156-13-22
523 Li, Y., Stocks, M., Hemmila, S., Kallman, T., Zhu, H., Zhou, Y., … Lascoux, M. (2010). Demographic histories
524   of four spruce (Picea) species of the Qinghai-Tibetan Plateau and neighboring areas inferred
525   from multiple nuclear loci. *Molecular Biology and Evolution*, *27*, 1001–1014.
526   https://doi.org/10.1093/molbev/msp301
527 Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., & Corander, J. (2016). Fundamentals and recent
528   developments in approximate Bayesian computation. *Systematic Biology*, syw077-syw077.
529   https://doi.org/10.1093/sysbio/syw077
530 Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., … Wang, J. (2014). Population genomics
531   reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, *157*, 785–794.
532   https://doi.org/10.1016/j.cell.2014.03.054
533 Lohse, K., Harrison, R. J., & Barton, N. H. (2011). A general method for calculating likelihoods under the
534   coalescent process. *Genetics*, *189*, 977–987. https://doi.org/10.1534/genetics.111.129569
535 Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T., & Murray, C. J. (2006). Global and regional burden
536   of disease and risk factors, 2001: systematic analysis of population health data. *Lancet*, *367*,
537   1747–1757.
538 Marin, J. M., Pudlo, P., Robert, C. P., & Ryder, R. J. (2012). Approximate Bayesian computational
539   methods. *Statistics and Computing*, *22*, 1167–1180. https://doi.org/10.1007/s11222-011-9288-2
540 McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical*
541   *Transactions of the Royal Society B: Biological Sciences*, *360*, 1387–1393.
542   https://doi.org/10.1098/rstb.2005.1673
543 Nadachowska-Brzyska, K., Burri, R., Olason, P. I., Kawakami, T., Smeds, L., & Ellegren, H. (2013).
544   Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-
545   genome re-sequencing data. *PLoS Genetics*, *9*, e1003942.
546   https://doi.org/10.1371/journal.pgen.1003942
547 Narum, S. R., Buerkle, C. A., Davey, J. W., Miller, M. R., & Hohenlohe, P. A. (2013). Genotyping-by-
548   sequencing in ecological and conservation genomics. *Molecular Ecology*, *22*, 2841–2847.
549   https://doi.org/10.1111/mec.12350

550 Pavlidis, P., Laurent, S., & Stephan, W. (2010). msABC: a modification of Hudson's ms to facilitate multi-
551   locus ABC analysis. *Molecular Ecology Resources*, *10*, 723–727. https://doi.org/10.1111/j.1755-
552   0998.2010.02832.x
553 Prangle, D. (2017). Adapting the ABC distance function. *Bayesian Analysis*, *12*, 289–309.
554   https://doi.org/10.1214/16-BA1002
555 Prangle, D., Blum, M. G. B., Popovic, G., & Sisson, S. A. (2014). Diagnostic tools for approximate Bayesian
556   computation using the coverage property. *Australian & New Zealand Journal of Statistics*, *56*,
557   309–329. https://doi.org/10.1111/anzs.12087
558 Quéméré, E., Amelot, X., Pierson, J., Crouau-Roy, B., & Chikhi, L. (2012). Genetic data suggest a natural
559   prehuman origin of open habitats in northern Madagascar and question the deforestation
560   narrative in this region. *Proceedings of the National Academy of Sciences of the United States of*
561   *America*, *109*, 13028–33. https://doi.org/10.1073/pnas.1200153109
562 R Core Team. (2016). R: A language and environment for statistical computing. *R Foundation for*
563   *Statistical Computing, Vienna, Austria.* Retrieved from https://www.R-project.org/
564 Robinson, J. D., Bunnefeld, L., Hearn, J., Stone, G. N., & Hickerson, M. J. (2014). ABC inference of multi-
565   population divergence with admixture from unphased population genomic data. *Molecular*
566   *Ecology*, *23*, 4458–4471. https://doi.org/10.1111/mec.12881
567 Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple
568   genome sequences. *Nature Genetics*, *46*, 919–925. https://doi.org/10.1038/ng.3015
569 Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history.
570   *Nature Reviews Genetics*, *16*, 727–740. https://doi.org/10.1038/nrg4005
571 Shafer, A. B. A., Gattepaille, L. M., Stewart, R. E. A., & Wolf, J. B. W. (2015). Demographic inferences
572   using short-read genomic data in an approximate Bayesian computation framework: In silico
573   evaluation of power, biases and proof of concept in Atlantic walrus. *Molecular Ecology*, *24*, 328–
574   345. https://doi.org/10.1111/mec.13034
575 Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2016).
576   Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic
577   inference. *Methods in Ecology and Evolution*, n/a-n/a. https://doi.org/10.1111/2041-210X.12700
578 Sousa, V. C., Beaumont, M. A., Fernandes, P., Coelho, M. M., & Chikhi, L. (2012). Population divergence
579   with or without admixture: selecting models using an ABC approach. *Heredity*, *108*, 521–530.
580   https://doi.org/10.1038/hdy.2011.116
581 Staab, P. R., Zhu, S., Metzler, D., & Lunter, G. (2015). scrm: efficiently simulating long sequences using
582   the approximated coalescent with recombination. *Bioinformatics (Oxford, England)*, *31*, 1680–2.
583   https://doi.org/10.1093/bioinformatics/btu861
584 Stocks, M., Siol, M., Lascoux, M., & De Mita, S. (2014). Amount of information needed for model choice
585   in Approximate Bayesian Computation. *PLoS ONE*, *9*, 1–13.
586   https://doi.org/10.1371/journal.pone.0099581
587 Sunnaker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M., & Dessimoz, C. (2013). Approximate
588   Bayesian Computation. *PLoS Computational Biology*, *9*.
589   https://doi.org/10.1371/journal.pcbi.1002803
590 Waples, R. S., & Do, C. (2008). LDNE : a program for estimating effective population size from data on
591   linkage disequilibrium. *Molecular Ecology Resources*, *8*, 753–756.
592   https://doi.org/10.1111/j.1755-0998.2007.02061.x
593 Wegmann, D., Leuenberger, C., Neuenschwander, S., & Excoffier, L. (2010). ABCtoolbox: a versatile
594   toolkit for approximate Bayesian computations. *BMC Bioinformatics*, *11*, 116–116.
595   https://doi.org/10.1186/1471-2105-11-116

596   Zinck, J. W. R., & Rajora, O. P. (2016). Post-glacial phylogeography and evolution of a wide-ranging
597   highly-exploited keystone forest tree, eastern white pine (Pinus strobus) in North America: single
598   refugium, multiple routes. *BMC Evolutionary Biology*, *16*, 56–56.
599   https://doi.org/10.1186/s12862-016-0624-1
600
601

602   **Data accessibility**
603   All relevant information to reproduce this study is included in this manuscript and supporting
604   information.
605
606

607   **Author contributions**
608   J.S.E and S.N.A conceived the study. J.S.E performed simulations and analysed the data. J.S.E wrote the
609   manuscript with the help of corrections and comments from S.N.A.
610

611   **Supporting information**
612   Additional supporting information including methods, figures and scripts can be found online.
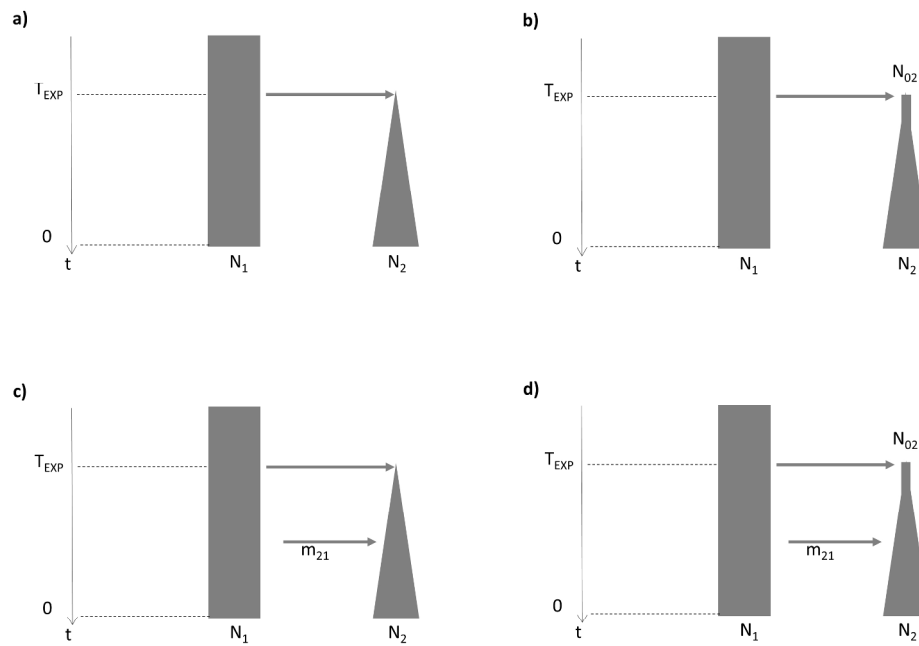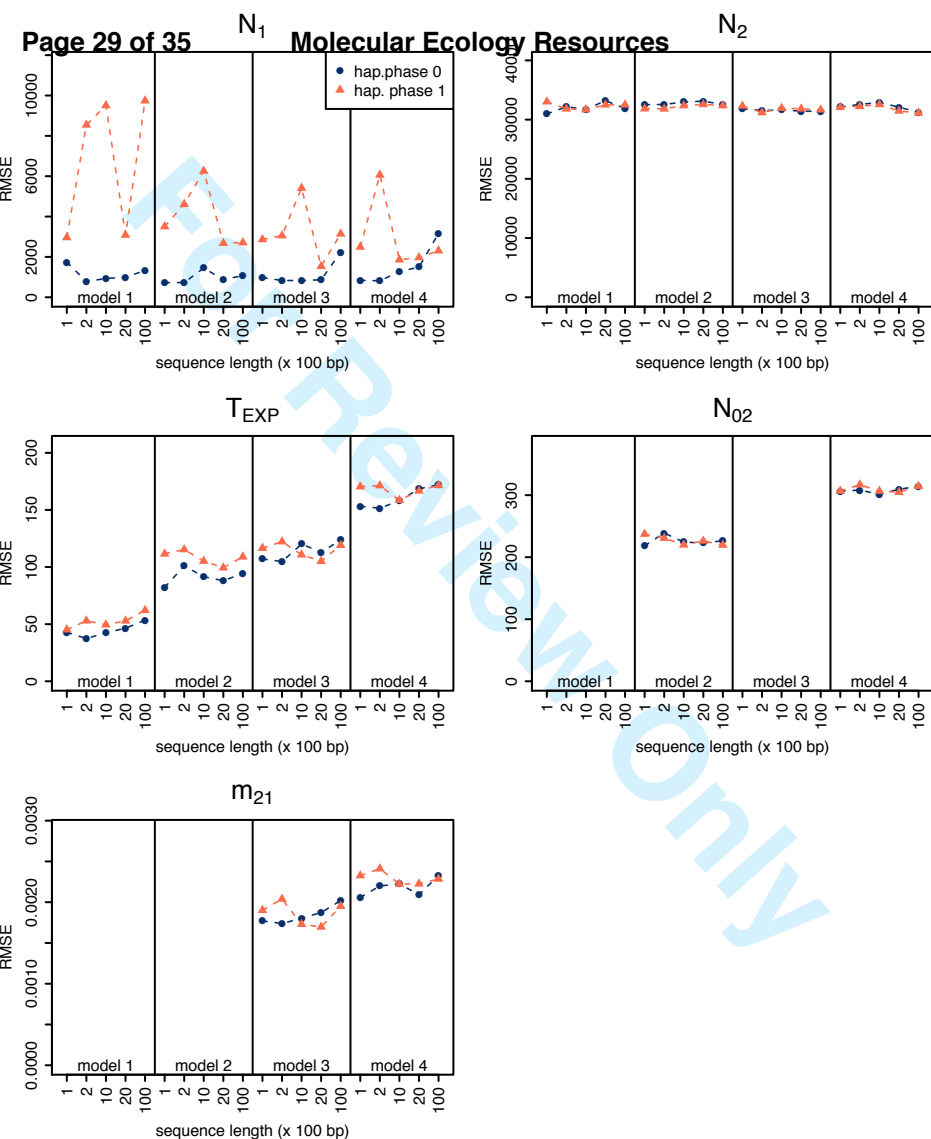613

Figure and table captions

Figure 1. Demographic models. a) Model 1: A three-parameter model of expansion featuring colonization of a 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}.

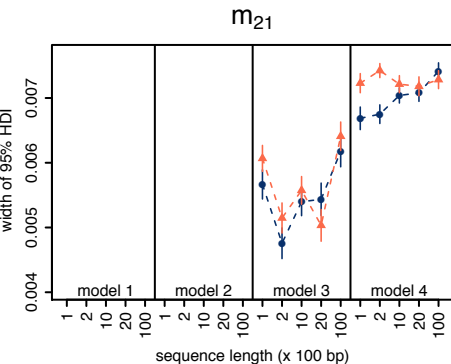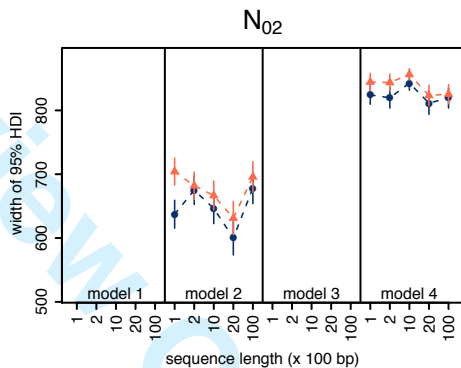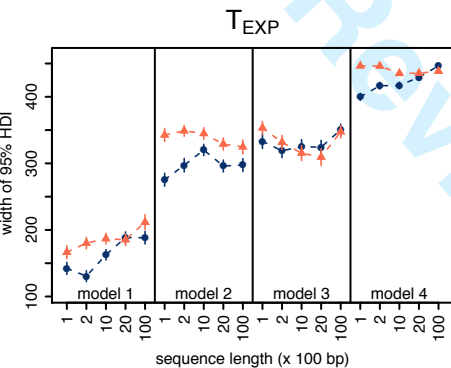**Figure 2.** RMSE calculated from the results of ABC analyses of 20 different combinations of demographic models and sampling designs (x-axis). For each combination, ABC was performed on simulated datasets summarized with statistics including linkage-based measures (hap. phase 1) and on the same set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

Figure 3. Width of the 95% highest posterior density values 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 a more detailed explanation.

Figure 4. Accuracy of model 1 estimates for current population sizes and time of expansion from population 1 to population 2 in generations. Points represent the mode of the posterior distribution for each POD simulated. Dotted lines represent the positions corresponding to a perfect estimation (posterior mode = true value). Error bars represent 95% HDI for each POD estimated. The top panel represents runs where haplotype phase is known, the bottom panel shows the result for the same dataset but unphased. **a)** Model 1 simulated with datasets of type 1 (10,000 sequences of 100bp). **b)** Model 1 simulated with datasets of type 5 (100 sequences of 10 kb).

Figure 5. RMSE of model parameters for different fixed values of TEXP. 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 plotted together with different characters and colours. Note that the TEXP values are represented on a log scale for better visibility on results for recent demographic events.

Figure 6. Relative prediction error (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 7. 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-

659 observed datasets for both the ABC and the SFS runs. HDI and CI widths were calculated from 100
660 bootstraps. PODs had 10,000 sequences of 100bp genotyped in 20 diploid individuals.

661

662

663 **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}	U(2:500)	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)	-

664

665

666

667 **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

668

Figure 1. Demographic models. a) Model 1: A three-parameter model of expansion featuring colonization of a 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}.

254x190mm (300 x 300 DPI)

b.