MOLECULAR ECOLOGY
RESOURCES

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

Title

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

Authors

Joane S. Elleouet, Sally N. Aitken

Address

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

Keywords

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

Corresponding author

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

Running title

Inferring recent population history with ABC

Abstract

Approximate Bayesian computation (ABC) is widely used to infer demographic history of populations and species using DNA markers. Genomic markers can now be developed for non-model species using a class of sequencing methods called reduced representation library (RRL), which selects a fraction of the genome using targetted sequence capture or restriction enzymes (Genotype-by-sequencing, GBS). The influence of the amount of information present in RRL datasets on the quality of demographic inference needs assessing. Here, we explored the influence of marker number and length and knowledge of the gametic phase on the quality of demographic inference performed with ABC. We focused on 2-population models of recent spatial expansion with varying number of unknown parameters. Performing ABC on simulated datasets with known parameter values, we found that the timing of a recent spatial expansion event can be precisely estimated in a 3-parameter model. Taking into account uncertainty in additional parameters such as initial population size and migration rate decreased the precision of the inference dramatically for all parameters except current population sizes. Phasing haplotypes did not improve results, regardless of sequence length, and numerous short sequences were as valuable as fewer, longer sequences. ABC results were similar to results obtained with an alternative method based on the site frequency spectrum (SFS) when performed with unphased GBS-type markers. We conclude that unphased GBS-type datasets can be sufficient to precisely infer simple demographic models, and discuss possible improvement for the use of phased haplotypes.

Introduction

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

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

Demographic inference in natural populations of non-model organisms

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

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

Approximate Bayesian computation (ABC) and other approaches

In this paper, our aim is to explore the ABC method for datasets obtained from reduced-representation library sequencing in non-model organisms. We also compare the obtained results with results from a SFS approach based on the approximation of the composite likelihood (Excoffier & Foll,

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

Previous work exploring the ABC method

The need to test the inference power of datasets for demographic models of interest has been recognized in recent years. Several studies show the use of preliminary simulations testing the number and length of markers and the number of individuals (Sousa et al., 2012; Stocks et al., 2014), the type of molecular markers (Cabrera & Palsbøll, 2017) and the choice of summary statistics and models considered (Benazzo et al., 2015; Guillemaud et al., 2010; S. Li & Jakobsson, 2012; Sousa et al., 2012; Stocks et al., 2014). As most scientists have switched to using genome-wide data, there is a need to expand this set of simulation studies to test and understand the power of different types of genomic data. As part of such an effort, Li & Jakobsson (2012) simulated large, phased genomic datasets comparable to human genomic datasets at the time. Under 2-population split models, they found that ABC produces accurate estimates for most but not all parameters and concluded ABC is well suited to large genomic datasets summarized with LD-based statistics. Robinson et al. (2014) tested the effects of the number and length of unphased genomic sequences and compared them to the effect of the number of individuals sequenced for the inference of three-population admixture models. They found that increasing the number and length of sequences was more beneficial than increasing sample size. Shafer et al. (2015) investigated the power of ABC on short diploid sequences obtained by GBS. They focused on a wide range of simple 1-population and 2-population models with bottleneck, growth, migration and a combination of these parameters. They found that population changes such as ancient temporary bottleneck would not be inferred correctly regardless of the number of markers available. This set of studies provides valuable information about the use of genomic data in ABC. Our aim is to extend this knowledge by directly comparing ABC results from molecular markers obtained with different types of RRL sequencing techniques and with different levels of genomic knowledge. This will hopefully help future ABC users who do not have access to complete genomic data to select methods

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

General model and datasets

Here, we focus on a set of 2-population models of demic expansion that are applicable to studies of species invasion, reintroduction, or natural colonization. We tested the power of ABC on these models using a range of marker sets obtainable by RRL methods: datasets with a large number of short genomic reads would correspond to single-end GBS sequencing, whereas fewer but longer diploid sequences correspond to a targeted enrichment approach. For each type of dataset, we quantified the potential benefits of knowing the gametic phase of sequence markers by including or excluding linkage-related statistics at the data-summarizing step. We expected to observe an improvement in the inference for datasets with long sequences. For each model assessed, we also tested the effect of time since the colonization event occurred. We hypothesize that recent events might be inferred more accurately with datasets containing linkage information, due to the generally higher rate of recombination compared to mutation, and to the potential information contained in long haplotypes. This part of the analysis is also motivated by the fact that overestimates of divergence times are a common result of demographic inference in empirical studies (Holliday et al., 2010) and this upward bias has been found for some demographic scenarios in simulation studies (Benazzo et al., 2015). We therefore aim to explore this potential bias by testing increasingly old events within the same models. Finally, we compared our ABC results with those obtained from an approximate likelihood method using the site frequency spectrum from simulated reduced-representation libraries. As they provide millions of genome-wide SNPs without ascertainment bias, restriction enzyme-based genomic sequencing techniques seem to be particularly

well suited to SFS-based inference methods. Comparing SFS results with ABC results on a range of models and datasets will inform future work on demographic inference in non-model organisms.

Methods

Demographic models

We focused on a basic 2-population model of demic expansion (fig.1a). A pre-existing population, population 1, is of constant size N_1. At time T_{EXP} before present, the spatial population expansion begins: population 2 is created by 2 migrants from population 1. Population 2 then grows exponentially between times T_{EXP} and T=0 (present) to size N_2. The rate of expansion is defined by the other parameter values through the formula $r=\log(\frac{N_{02}}{N_2})/T_{EXP}$. Model 1 therefore has just 3 independent unknown parameters: N_1, N_2, and T_{EXP}. We created additional models of increasing complexity by adding parameters. In models 2 and 4, the number of founders of population 2, N_{02}, is unknown (fig.1b and fig.1d); in models 3 and 4, migration is allowed from population 1 to population 2, with the parameter m_{21} describing a per-generation migration rate (fig.1c and fig.1d). In all four models described above, mutation and recombination rate are fixed. Parameter priors were chosen to approximate studies of postglacial expansion of temperate tree species (Table 1); population sizes are generally large and expansion time has an upper limit of 500 generations. However, we set priors with wide ranges so that results would inform students of other biological systems.

Genomic datasets simulated

For each of the four models, we created a set of 1 million simulations with each of the five types of datasets described below, with a fixed number of 10 diploid individuals sampled per population. For

datasets corresponding to single-end RADseq sequencing techniques, we simulated 10,000 independent DNA sequences of 100bp each. For datasets corresponding to sequence capture methods, we created 100 independent DNA sequences of 10kb each. Additionally, we explored a range of possible configurations between these two types of datasets (table 2). With 4 models and 5 types of datasets, we obtained a total of 20 combinations of models and datasets, each with a million simulations. We used the program scrm (Staab et al., 2015) in R (R Core Team, 2016) to compute the simulations, using custom scripts inspired by scripts from Shafer et al. (2015) and available in the supporting information.

Summary statistics

For each simulated dataset we used the program msABC (Pavlidis et al., 2010) to compute all summary statistics available, including diversity statistics (number of segregating sites and θ estimates), summaries of the SFS (Tajima's D and Fay and Wu's H), differentiation measures (pairwise F_{ST}, number of private and shared polymorphisms), the Thomson estimator of T_{MRCA}, and its variance. To test the effect that knowing haplotype information has on inference, the ABC analysis was performed twice on each model-dataset type combination. The first time, we summarized data using only the statistics mentioned above, which are calculated at the SNP level and therefore are available when the gametic phase of the diploid sequences is unknown. The second inference was performed on the same dataset, but additional statistics based on linkage information were used to summarize the data: Zns (Kelly, 1997), dvk and dvh (Depaulis & Veuille, 1998). These additional statistics are calculated at the haplotype level, they are therefore only available in cases where the gametic phase of the diploid sequences is known. For each set of simulations, we computed the mean and variance of every statistic over all sequence markers in the dataset. As a result, 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

the prior ranges, and parameters were estimated for each of them with unchanged settings. For each parameter, we calculated the root mean squared error (RMSE) from the results of these i=1000 random simulations, following equation (1):

$$(1)\ \mathcal{E} = \sqrt{\frac{\sum_1^i (\hat{\theta} - \theta^*)^2}{i}}$$

where θ^* is the parameter value used for the simulation and $\hat{\theta}$ is the estimate defined as the parameter value with highest posterior probability.

For comparison purposes, we computed a complementary measure of precision and accuracy called the relative prediction error (RPE), the ratio of the variance of the posterior over the variance of the prior, which follows equation (2):

$$(2)\ \mathcal{E} = \frac{\sum_1^i (\hat{\theta} - \theta^*)^2}{Var(\theta^*)} \times \frac{1}{i}$$

where Var(θ^*) is the variance of the prior distribution. The advantage of using RPE as a validation statistic is that it directly indicates the contribution of the genetic dataset to the estimation of the posterior. Another attractive feature of the RPE is that it allows comparisons between parameters, as it scales from 0 (precise estimate) to 1 and beyond (in the case of a consistent bias in estimation).

The 95% highest posterior density interval (HDI) was calculated as an additional measure of precision. It is defined as the shortest continuous interval with an integrated posterior density of a certain value (Wegmann et al., 2010). As such, it is an ABC equivalent of the Bayesian credible interval. The 95% HDI was calculated on 100 PODs for each set of 1M simulations.

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

To test the effect of the time of expansion on the precision of the ABC estimation, we created 100 PODs for each set of 1M simulations and each of the following T_{EXP} values: 2, 5, 10, 20, 50, 75, 100, 150, 200, 300, 400 and 500 generations. RMSE and 95%HDI were calculated from the results of each set of 100 PODs.

Comparing ABC and SFS estimation

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

Results

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

Effect of model complexity on the precision of parameter estimates

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

Not all parameters were sensitive to the addition of parameters in the models: the precision of contemporary population size estimates N_1 and N_2 was independent of model complexity. RMSE values for N_1, which were set as constant over generations, were below 10,000 for the four models assessed (fig.2). 95% highest posterior density intervals ranged from 1000 to 4000. For N_2, the contemporary population 2 size after exponential growth, 95% HDI intervals were about as wide as the prior range, indicating a failure to estimate this parameter in all four models.

The expansion time T_{EXP} was generally well estimated in model 1, which is the simplest, 3-parameter model (fig.2) with no migration between demes and the number of founders set to 2. For this model, 95% HDI intervals show moderate precision with ranges as wide as 150 to 200 generations. The precision of the T_{EXP} estimation was mediocre for more models where the number of founders N_{02} (model 2) is unknown and needs to be estimated, or where migration from population 1 to population 2 is likely (model 3). For these two models, RMSE values of T_{EXP} were around 110, and 95% HDI were as wide as 300 generations. The ABC analysis on the 5-parameter model (model 4) was unable to recover the true T_{EXP} value (95% HDI ≥ 400).

Estimates of the number of founders of population 2 (N_{02}) and migration rate from population 1 to 2 (m_{21}) were also impacted by model complexity. In 4-parameter models, RMSE values for N_{02} were

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 recovered in the most complex model 4, and m_{21} RMSE values were high. Generally, the 95% highest posterior density intervals followed the same trends as the RMSE for N_{02} and m_{21} (fig.3): N_{02} and m_{21} were each impossible to infer in the presence of the other.

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

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

The addition of the linkage statistics available in msABC brought no notable improvement in the RMSE and 95% HDI of parameter estimates for all models (fig.2 and fig.3). It even seems to make the estimation of N_1 less precise in model 1, as shown by an increase of up to 36000 of 95% HDI values (fig.3). ABC performance on models 3 and 4 seemed to be slightly more dependent on sequence length, with the inference on large sequences marginally benefitting from haplotype information.

Quality of parameter estimates across prior ranges

For each parameter, we visualized estimated values and 95% HDI of ABC results in relation to true parameter values to assess performance over the prior range. Results for the 3-parameter model (model 1) and dataset types 1 and 5 are shown in fig.4a and fig.4b, respectively. Results for the complete set of models are available in supplemental fig.S1. Consistently across models, estimates of N_2 are largely inaccurate regardless of the true value, with HDI ranges as wide as the prior range. Conversely, N_1 estimates are accurate in all models regardless of the true N_1 value. Unlike N_1 and N_2, the values of T_{EXP}, N_{02} and m_{02} do have an impact on the precision of their respective estimates. Accuracy and precision of T_{EXP} estimates for models 1 to 3 decrease with increasing true value. N_{02}, the number of founders of population 2, is estimated in model 2 (4 parameters) and model 4 (5 parameters). Inferences on model 4 failed to estimate N_{02} across all dataset types and haplophases. For inference of model 2, estimates of N_{02} values below 150 are fairly accurate, with 95% HDI intervals as low as 20 when simulating datasets with marker lengths 100 and 200. With datasets of fewer, longer sequences, a smaller range of low N_{02} true values lead to a high level of precision. Finally, the per-generation migration rate from population 1 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} values are, the less accurate and precise their estimates become (fig.S1). m_{21} values below 0.002 could be estimated with moderate precision in model 3, and very imprecisely in model 4 (fig.S1).

Effect of the time of the expansion event on the estimation

We tested whether older expansion events are generally more difficult to characterize than recent ones within the time range specified by the prior. To do this, we studied the effect of the true T_{EXP} value on the precision of parameter estimates. We find different trends foramong the 4 models (fig.5, S2, and S3). The precision of inference on model 1 is high at low T_{EXP} values and decreases as T_{EXP}

increases. For model 2, older events seem to be generally better inferred. However, the precision of T_{EXP} estimantes decreases with T_{EXP} but the precision of N_{02} has the opposite trend, suggesting that estimating both T_{EXP} and N_{02} precisely is almost impossible. For both model 1 and 2, there is a decrease in RMSE and 95% HDI estimates at high T_{EXP} values, but this probably only reflects the fact that prediction values are bounded to the upper prior limit (fig.S2 and S3). Model 3 shows best results for moderately recent expansion events (20-75 generations), as shown by estimates of T_{EXP} and m_{21}. Finally, results for model 4 show high values of RMSE and 95% HDI for all parameters, but it seems that the precision of T_{EXP}, N_{02}, and m_{21} is generally higher for older expansion events.

Looking at results for individual parameters across all models, we note that the estimation of T_{EXP} is better for a more recent expansion event as long as there is no migration between populations (fig.5). The precision of N_{02} estimates is highly dependent on T_{EXP}, especially for model 2, where RMSE values decrease from 400 to as low as 100 as T_{EXP} increases. Estimation of the migration parameter m_{21} also seem to be affected by the timing of population expansion but the relationship is complex in model 3. In model 4, RMSE values for m_{21} decrease as T_{EXP} increases.

Comparing ABC with SFS estimation using an approximate composite likelihood

Figures 6 and 7 illustrate the performance of ABC and approximate composite likelihood from the SFS for all models performed with datasets of 10,000 100-kb sequences. Both methods gave similar results in terms of precision of parameter estimates. The SFS-based method performed slightly better than ABC in the model with migration (model 3), although this superiority is not observed in the width of confidence intervals. ABC resulted in a lower relative prediction error for model 2 in the estimation of N_{02} (fig.6), but again the width of confidence intervals was similar across methods.

Discussion

We explored the ability of approximate Bayesian computation to characterize a recent event of spatial expansion from one population of constant size to a new growing population, a model which can be broadly applied to studies of species range expansion, invasion biology, or reintroduction of endangered species. Our results show that higher model complexity provided results of lower quality. Precisely dating a spatial population expansion event that occurred in the recent past was not possible for complex models (more than 4 parameters). The simplest model with only three parameters (N_1, N_2 and T_{EXP}) could be partially recovered with a precise estimate of timing of expansion and size of source population (N_1, assuming it is constant) but a poor estimate of the size of the growing, newly founded population (N_2). Failure to estimate N_2 does not come as a surprise: estimates of past changes in effective population size from one punctual sampling event commonly rely on linkage information between markers, a calculation not readily available in ABC packages (Beaumont, 2003; Waples & Do, 2008). The inference of models including the number of founders of population 2 (N_{02}) and/or per-generation migration rate from population 1 to population 2 (m_{21}) was only partially successful. The number of founders and time of founding could not be jointly estimated precisely. We found that the timing of a demographic event has an effect on the precision of parameter estimates with ABC: the time of expansion itself was more precisely estimated when it was moderately recent, a trend also observed in Li and Jakobsson (2012). The presence of migration makes inference of the time of expansion impossible when it is more recent than about 10 generations, and historical events could only be inferred for a narrow time range (20 to 50 generations ago). The difficulty of estimating the time of a

population founding event with subsequent migration has also been reported in Robinson et al. (2014). The number of founders of population 2 and the per-generation migration rate from population 1 to population 2 were both estimated more precisely for events that are more ancient.

Implications of including haplotype information

Analyses based on unphased sequences exploring similar models as here has shown encouraging results (Robinson et al., 2014). However, no study to date has explicitly compared datasets of phased and unphased sequences using the same models and same amounts of data. Here, we quantified the benefits of using haplotypes of phased sequences over single SNPs by including or leaving out LD-based and haplotype-level statistics at the data summarization step of the ABC inference. Adding haplotype information to simulated datasets made surprisingly little difference compared to identical ABC runs with phased data. The reason why phasing the data did not improve inference could be that the extent of linkage the chosen statistics are sensitive to is different from the linkage actually present in the simulated data. In the present study, we decided to use all statistics available in msABC. We chose to simulate sequences with a fixed per-nucleotide recombination rate r of 10^{-8}, a value compatible with estimates in mammal species but much lower than estimates in other organisms such as plants. We simulated sequences spanning a maximum of 10^4 bp. This results in a maximum per-marker recombination rate of 10^{-4} per generation. Considering that the effective population sizes simulated were between 10^4 and 10^5, the population-wise per-generation recombination rate for a 10kb marker $4N_e r$ could be as high as 40 per generation. For a demographic event of at least a few generations, recombination rates of this order of magnitude should at least affect the variance of LD-based summary 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

of millions of simulations that correspond closely to an observed dataset requires approximating the distribution of sequence lengths, and this may also affect inferences, especially if the variance of summary statistics are included at the data summarization step.

Data structure and quantity

We kept constant the total amount of sequenced data over all five types of datasets. Our aim was to focus on other, less studied aspects of data collection design. Studies exploring the effects of the amount of data in ABC inference have generally found that increasing the number of genetic markers is more efficient than increasing the number of individuals sampled (Robinson et al., 2014; Stocks et al., 2014). These studies found a threshold number of markers beyond which the precision of parameter estimates and of model choice reach a plateau. Both Robinson et al (2014) and Li & Jakobsson (2012) have concluded 2000 loci is an optimal number, although this would depend on many other variables such as locus length, model complexity, and even true value of parameters (Shafer et al., 2015).

Comparing ABC to other methods

We did not find large differences in the precision of parameter estimates between ABC and the SFS-based likelihood method implemented in fastsimcoal 2. Shafer et al. (2015) found a similar result while comparing the performance of ABC with a SFS-based inference implemented in *δaδi* (Gutenkunst et al., 2009). They found that *δaδi* tends to overestimate the time of population split and bottleneck events, a trend not supported by our findings with *fastsimcoal*. In addition to parameter estimation, Shafer *et al.* (2015) tested the performance of both methods for model selection and found ABC more accurate, especially in the case of bottleneck scenarios.

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

456 ## Acknowledgements

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

460
461
462 ## References

463 Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored
464 populations. *Genetics*, *164*, 1139–1160.
465 Benazzo, A., Ghirotto, S., Vilaça, S. T., & Hoban, S. (2015). Using ABC and microsatellite data to detect
466 multiple introductions of invasive species from a single source. *Heredity*, *115*, 262–272.
467 https://doi.org/10.1038/hdy.2015.38
468 Bertorelle, G., Benazzo, A., & Mona, S. (2010). ABC as a flexible framework to estimate demography over
469 space and time: Some cons, many pros. *Molecular Ecology*, *19*, 2609–2625.
470 https://doi.org/10.1111/j.1365-294X.2010.04690.x
471 Blum, M. G. B., Nunes, M. A., Prangle, D., & Sisson, S. A. (2013). A comparative review of dimension
472 reduction methods in approximate Bayesian computation. *Statistical Science*, *28*, 189–208.
473 https://doi.org/10.1214/12-STS406
474 Boitard, S., Rodríguez, W., Jay, F., Mona, S., & Austerlitz, F. (2016). Inferring population size history from
475 large samples of genome-wide molecular data: an approximate Bayesian computation approach.
476 *PLOS Genetics*, *12*, e1005877–e1005877. https://doi.org/10.1371/journal.pgen.1005877
477 Cabrera, A. A., & Palsbøll, P. J. (2017). Inferring past demographic changes from contemporary genetic
478 data: A simulation-based evaluation of the ABC methods implemented in DIYABC. *Molecular*
479 *Ecology Resources*. https://doi.org/10.1111/1755-0998.12696
480 Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature*, *325*,
481 31–36. https://doi.org/10.1038/325031a0
482 Chan, Y. L., Schanzenbach, D., & Hickerson, M. J. (2014). Detecting concerted demographic response
483 across community assemblages using hierarchical approximate Bayesian computation. *Molecular*
484 *Biology and Evolution*, *31*, 2501–15. https://doi.org/10.1093/molbev/msu187
485 Condon, E., & Cukier, M. (2016). Using Approximate Bayesian Computation to Empirically Test Email
486 Malware Propagation Models Relevant to Common Intervention Actions. In *2016 IEEE 27th*
487 *International Symposium on Software Reliability Engineering (ISSRE)* (pp. 287–297). IEEE.
488 https://doi.org/10.1109/ISSRE.2016.24
489 Csilléry, K., Blum, M. G. B., Gaggiotti, O. E., & François, O. (2010). Approximate Bayesian Computation
490 (ABC) in practice. *Trends in Ecology & Evolution*, *25*, 410–8.
491 https://doi.org/10.1016/j.tree.2010.04.001
492 Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-
493 wide genetic marker discovery and genotyping using next-generation sequencing. *Nature*
494 *Reviews Genetics*, *12*, 499–510. https://doi.org/10.1038/nrg3012
495 Depaulis, F., & Veuille, M. (1998). Neutrality tests based on the distribution of haplotypes under an
496 infinite-site model. *Molecular Biology and Evolution*, *15*, 1788–1790.
497 https://doi.org/10.1093/oxfordjournals.molbev.a025905
498 Dussex, N., Wegmann, D., & Robertson, B. C. (2014). Postglacial expansion and not human influence best
499 explains the population structure in the endangered kea (Nestor notabilis). *Molecular Ecology*,
500 *23*, 2193–2209. https://doi.org/10.1111/mec.12729
501 Excoffier, L., & Foll, M. (2011). fastsimcoal: a continuous-time coalescent simulator of genomic diversity
502 under arbitrarily complex evolutionary scenarios. *Bioinformatics*, *27*, 1332–1334.
503 https://doi.org/10.1093/bioinformatics/btr124

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Prangle, D. (2017). Adapting the ABC distance function. *Bayesian Analysis*, *12*, 289–309. https://doi.org/10.1214/16-BA1002

Prangle, D., Blum, M. G. B., Popovic, G., & Sisson, S. A. (2014). Diagnostic tools for approximate Bayesian computation using the coverage property. *Australian & New Zealand Journal of Statistics*, *56*, 309–329. https://doi.org/10.1111/anzs.12087

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

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

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

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

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

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

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

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

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

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

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

Waples, R. S., & Do, C. (2008). LDNE : a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, *8*, 753–756. https://doi.org/10.1111/j.1755-0998.2007.02061.x

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

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

Data accessibility

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

Author contributions

J.S.E and S.N.A conceived the study. J.S.E performed simulations and analysed the data. J.S.E wrote the manuscript with the help of corrections and comments from S.N.A.

Supporting information

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

Figure and table captions

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

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

Table 1. Model parameters with their associated prior ranges

estimated in models	Parameter	Symbol	Prior range	Unit
-	Mutation rate	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)	-

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

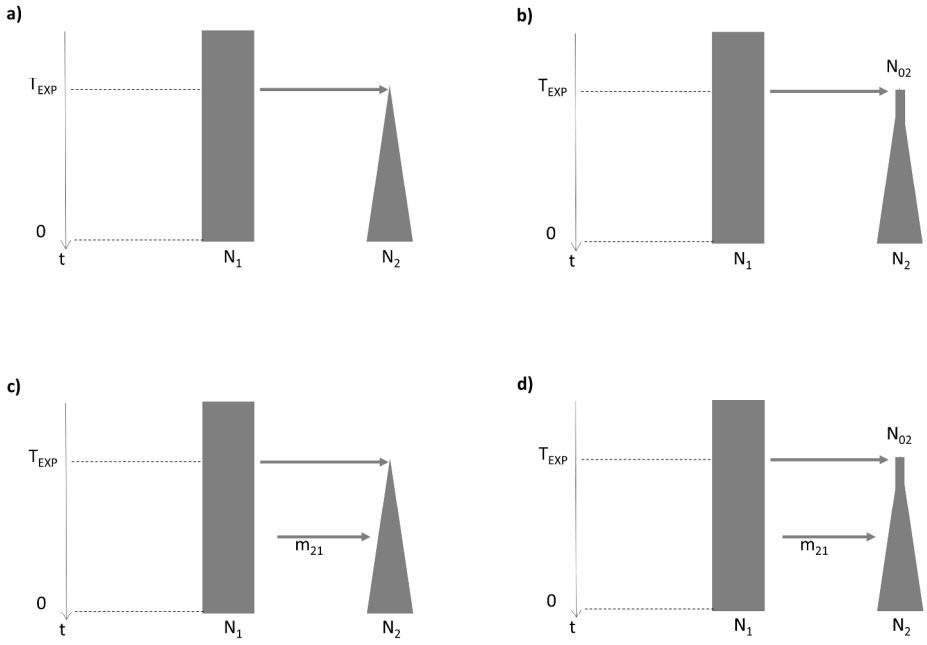

Figure 1. Demographic models. a) Model 1: A three-parameter model of expansion featuring colonization of a new population 2 by 2 diploid individuals from population 1 at time T_{EXP}. Population 1 is of constant size N_1, whereas population 2 grows exponentially to size N_2, its size at present. b) Model 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)