Maximum likelihood inference of population size contractions from microsatellite data

Raphaël Leblois\textsuperscript{a,b,f,*}, Pierre Pudlo\textsuperscript{a,d,f}, Joseph Néron\textsuperscript{b}, François Bertaux\textsuperscript{b,e}, Champak Reddy Beeravolu\textsuperscript{a}, Renaud Vitalis\textsuperscript{a,f}, François Rouset\textsuperscript{c,d}

\textsuperscript{a}INRA, UMR 1062 CBGP (INRA-IRD-CIRAD-Montpellier Supagro), Montpellier, France
\textsuperscript{b}Muséum National d’Histoire Naturelle, CNRS, UMR OSEB, Paris, France
\textsuperscript{c}Université Montpellier 2, CNRS, UMR ISEM, Montpellier, France
\textsuperscript{d}Université Montpellier 2, CNRS, UMR I3M, Montpellier, France
\textsuperscript{e}INRIA Paris-Rocquencourt, BANG team, Le Chesnay, France
\textsuperscript{f}Institut de Biologie Computationnelle, Montpellier, France

Abstract

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology. In the present work, we develop a maximum likelihood method for the inference of past changes in population size from microsatellite allelic data. Our method is based on importance sampling of gene genealogies, extended for new mutation models, notably the generalized stepwise mutation model (GSM). Using simulations, we test its performance to detect and characterize past reductions in population size. First, we test the estimation precision and confidence intervals coverage properties under ideal conditions, then we compare the accuracy of the estimation with another available method (\textsc{MsVar}) and we finally test its robustness to misspecification of the mutational model and population structure. We show that our method is very competitive compared to alternative ones. Moreover, our implementation of a GSM allows more accurate analysis of microsatellite data, as we show that violations of a single step mutation assumption induce very high bias towards false bottleneck detection rates. However, our simulation tests also showed some important limits, which most importantly are large computation times.

\*Corresponding author
Email address: raphael.leblois@supagro.inra.fr (Raphaël Leblois)
for strong disequilibrium scenarios and a strong influence of some form of unaccounted population structure. This inference method is available in the latest implementation of the MIGRAINE software package.

*Keywords:* demographic inference, maximum likelihood, coalescent, importance sampling, microsatellites, bottleneck, population structure, mutation processes
1. Introduction

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology, e.g. for understanding the effects of environmental changes on the distribution of organisms. From a conservation perspective, a severe reduction in population size, often referred to as a “population bottleneck”, increases rate of inbreeding, loss of genetic variation, fixation of deleterious alleles, and thereby greatly reduces adaptive potential and increases the risk of extinction (Lande, 1988; Frankham et al., 2006; Keller and Waller, 2002; Reusch and Wood, 2007). However, characterizing the demographic history of a species with direct demographic approaches requires the monitoring of census data, which can be extremely difficult and time consuming (Williams, Nichols and Conroy, 2002; Schwartz, Luikart and Waples, 2007; Bonebrake et al., 2010). Moreover, direct approaches cannot give information about past demography from present-time data. A powerful alternative relies on population genetic approaches, which allow inferences on the past demography from the observed present distribution of genetic polymorphism in natural populations (Schwartz, Luikart and Waples, 2007; Lawton-Rauh, 2008).

Until recently, most indirect methods were based on testing whether a given summary statistic (computed from genetic data) deviates from its expected value under an equilibrium demographic model (Cornuet and Luikart, 1996; Schneider and Excoffier, 1999; Garza and Williamson, 2001). Because of their simplicity, these methods have been widely used (see, e.g. Comps et al., 2001; Colautti et al., 2005, and the reviews of Spencer, Neigel and Leberg, 2000 and Peery et al., 2012). But they neither estimate the severity of the bottleneck nor its age or duration.

Although much more mathematically difficult and computationally demanding, likelihood-based methods outperform these moment-based methods by considering all available information in the genetic data (see Felsenstein, 1992; Griffiths and Tavaré, 1994; Emerson, Paradis and Thébaud, 2001, and the review of Marjoram and Tavaré, 2006). Among others, the software package MsVar (Beaumont, 1999; Storz and Beaumont, 2002) has been increasingly used to infer past demographic changes. MsVar assumes a demographic model consisting of a single isolated population, which has undergone a change in effective population size at some time in the past. It is dedicated to the analysis of microsatellite loci that
are assumed to follow a strict stepwise mutation model (SMM, Olha and Kimura, 1973).

In a recent study, Girod et al. (2011) evaluated the performance of \texttt{MsVar} by simulation. They have shown that \texttt{MsVar} clearly outperforms moment-based methods to detect past changes in population sizes, but appears only moderately robust to mis-specification of the mutational model: deviations from the SMM often induce “false” bottleneck detections on simulated samples from populations at equilibrium. Chikhi et al. (2010) also found a strong confounding effect of population structure on bottleneck detection using \texttt{MsVar}.

Thus, departures from the mutational and demographic assumptions of the model appear to complicate the inference of past population size changes from genetic data.

The present work extends the importance sampling (IS) class of algorithms (Stephens and Donnelly, 2000; de Iorio and Griffiths, 2004a,b) to coalescent-based models of a single isolated population with past changes in population size. Moreover, in the spirit of de Iorio et al. (2005), we also provide explicit formula for a generalized stepwise mutation model (GSM, Pritchard et al., 1999).

We have conducted three simulation studies to test the efficiency of our methodology on past contractions (i.e., bottlenecks) and its robustness against mis-specifications of the model. The first study aims at showing the ability of the algorithm to detect bottlenecks and to recover the parameters of the model (i.e., the severity of the size change and its age) on a wide range of bottleneck scenarios. In a second study, we compared the accuracy of our IS implementation with the MCMC approach implemented in \texttt{MsVar}. The third study tests the robustness of our method against mis-specification of the mutation model, and against the existence of a population structure not considered in the model. All analyses in these studies were performed using the latest implementation of the \texttt{Migraine} software package, available at the web page kimura.univ-montp2.fr/~rousset/Migraine.htm.

2. New approaches

Our goal is to obtain maximum likelihood (ML) estimates for single population models with a past variation in population size, described in Subsection 2.1. To this end, we describe the successive steps of the inference algorithm (Subsections 2.2, 2.3).
2.1. Demographic model

Figure 1: Representation of the demographic model used in the study.

$N$s are population sizes, $T$ is the time measured in generation since present and $\mu$ the mutation rate of the marker used. Those four parameters are the canonical parameters of the model. $\theta$s and $D$ are the inferred scaled parameters.

We consider a single isolated population with past size changes (Fig. 1). We denote by $N(t)$ the population size, expressed as the number of genes, $t$ generations away from the sampling time $t = 0$. Population size at sampling time is $N \equiv N(0)$. Then, going backward in time, the population size changes according to a deterministic exponential function until reaching an ancestral population size $N_{\text{anc}}$ at time $t = T$. Then, $N(t)$ remains constant, equal to $N_{\text{anc}}$ for all $t > T$. More precisely,

$$N(t) = \begin{cases} N \left( \frac{N_{\text{anc}}}{N} \right)^{\frac{T}{t}}, & \text{if } 0 < t < T, \\ N_{\text{anc}}, & \text{if } t > T. \end{cases} \quad (1)$$

To ensure identifiability, the parameters of interest are scaled as $\theta \equiv 2N\mu$, $\theta_{\text{anc}} \equiv 2\mu N_{\text{anc}}$ and $D \equiv T/2N$, where $\mu$ is the mutation rate per locus per generation. We often are interested in an extra composite parameter $N_{\text{ratio}} = \theta/\theta_{\text{anc}}$, which is useful to characterize the strength of the bottleneck. Finally, we also consider an alternative parametrization of the model using $\theta$, $\theta_{\text{anc}}$ and $D' \equiv \mu T$ in a few situations, for comparison between these two possible parameterizations.

2.2. Computation of coalescent-based likelihood with importance sampling

Because the precise genetic history of the sample is not observed, the coalescent-based likelihood at a given point of the parameter space is an integral over all possible histories,
i.e. genealogies with mutations, leading to the present genetic data. Following Stephens and Donnelly (2000) and de Iorio and Griffiths (2004a), the Monte Carlo scheme computing this integral is based here on importance sampling. The set of possible past histories is explored via an importance distribution depending on the demographical scenario and on the parameter values we are currently focused on. The best proposal distribution to sample from is the importance distribution leading to a zero variance estimate of the likelihood. Here it amounts to the model-based distribution of gene history conditioned by the present genetic data, which corresponds to all backward transition rates between successive states of the histories. As computation of these backward transition rates is often too difficult, we substitute this conditional distribution with an importance distribution, and introduce a weight to correct the discrepancy. Like the best proposal distribution, the actual importance distribution is a process describing changes in the ancestral sample configuration backward in time using absorbing Markov chains but do not lead to a zero variance estimate of the likelihood. Better efficiency of the importance sampling proposals allows to accurately estimate likelihoods by considering less histories for a given parameter value. Stephens and Donnelly (2000), de Iorio and Griffiths (2004a,b) and de Iorio et al. (2005) suggested efficient approximations that are easily computable. However the efficiency of the importance distribution depends heavily on the demographic model and the current parameter value.

The first main difference between our algorithm and those described in de Iorio and Griffiths (2004a,b) is the time inhomogeneity induced by the disequilibrium of our demographic model. Demographic models considered in the above cited literature and in Rousset and Leblois (2007, 2012), do not include indeed any change in population sizes. To relax the assumption of time homogeneity in (de Iorio and Griffiths, 2004b), we modify their equations (see Tables 1 and 2 of de Iorio and Griffiths, 2004b), so that all quantities depending on the relative population sizes now vary over time because of the population size changes. Thus, we must keep track of time in the algorithm to assign the adequate value to those time dependent quantities. To see how this is done, consider that the genealogy has been constructed until time $T_k$, and that, at this date, $n$ ancestral lineages remain. Under the importance distribution, the occurrence rate of a mutation event is then $n\theta$, and the occurrence rate of a coalescence event is $n(n-1)\lambda(t)$, where $\lambda(t) = N/N(t)$ is
the population size function introducing the disequilibrium. $\lambda(t)$ corresponds to parameter $1/q$ in de Iorio and Griffiths (2004b). The total jump rate at time $t \geq T_k$ is then

$$\Gamma(t) = n \left( (n - 1)\lambda(t) + \theta \right)$$

and the next event of the genealogy occurs at time $T_{k+1}$ whose distribution has density

$$\hat{P}(T_{k+1} \in dt) = \Gamma(t) \exp \left( - \int_{T_k}^t \Gamma(u) du \right) \, dt \quad \text{if } t \geq T_k.$$ 

Apart from these modifications, the outline of the IS scheme from de Iorio and Griffiths (2004b) is preserved (see section A.1 in the supplementary materials for more details).

We also develop specific algorithms to analyze data under the generalized stepwise mutation model (GSM), with infinite or finite number of alleles. This more realistic mutation model considers that multistep mutations occur and the number of steps involved for each mutation can be modeled using a geometric distribution with parameter $p$. The original algorithm of Stephens and Donnelly (2000) covers any finite mutation model but requires numerical matrix inversions to solve a system of linear equations, (see, e.g., Eqs (18) and (19) in Stephens and Donnelly, 2000). Time inhomogeneity requires matrix inversions each time the genealogy is updated by the IS algorithm. To bypass this difficulty, de Iorio et al. (2005) have successfully replaced the matrix inversions with Fourier analysis when considering a SMM with an infinite allele range. We extended this Fourier analysis in the case of a GSM with an infinite allele range. However, contrarily to the SMM, the result of the Fourier analysis for the GSM is a very poor approximation for cases with a finite range of allelic state as soon as $p$ is not very small (e.g. $< 0.1$). To consider a more realistic GSM with allele ranges of finite size, we propose to compute the relevant matrix inversions using a numerical decomposition in eigenvectors and eigenvalues of the mutation process matrix, $P$. Because the mutation model is not time-depend, this last decomposition is performed only once for a given matrix $P$. See A.4 for details about the GSM implementation.

Finally, several approximations of the likelihood, using products of approximate conditional likelihoods (PAC, Cornuet and Beaumont, 2007) and analytical computation of the probability of the last pair of genes, have been successfully tested to speed up computation times (see section A.2 in the supplementary materials).
2.3. Inference method

Following Rousset and Leblois (2007, 2012), we first define a set of parameter points via a stratified random sample on the range of parameters provided by the user. Then, at each parameter point, the multilocus likelihood is the product of the likelihoods for each locus, which are estimated via the IS algorithm described above. The likelihood inferred at the different parameter point is then smoothed by a Kriging scheme (Cressie, 1993). After a first analysis of the smoothed likelihood surface, the algorithm can be repeated a second time to increase the density of the grid in the neighborhood of a first maximum likelihood estimate. Finally, one- and two-dimensional profile likelihood ratios are computed, to obtain confidence intervals and graphical outputs (e.g. Fig. 2). Section A.3 in the supplementary materials explains how we tuned the parameters of the algorithm, namely the range of parameters, the size of parameter points and the number of genealogical histories explored by the IS algorithm.

A genuine issue, when facing genetic data, is to test whether the sampled population has undergone size changes or not. Thus, we derived a statistical test from the methodology presented above. It aims at testing between the null hypothesis that no size change occurred (i.e., \( N = N_{\text{anc}} \)) and alternatives such as a population decline or expansion (i.e., \( N \neq N_{\text{anc}} \)). At level \( \alpha \), our test rejects the null hypothesis if and only if 1 lies outside the \( 1 - \alpha \) confidence interval of the ratio \( N_{\text{ratio}} = N/N_{\text{anc}} \).

All those developments are implemented in the MIGRAINE software package. A detailed presentation of the simulation settings and validation procedures used to test the precision and robustness of the method are given in Section 5.

3. Results

3.1. Two contrasting examples

We begin with two contrasting simulated examples presented on Figs. 2a and 2b. The first one, corresponding to our baseline simulation (\( \theta = 0.4, D = 1.25 \) and \( \theta_{\text{anc}} = 40.0 \), case [0]), is an ideal situation in which the inference algorithm performs well due to the large amount of information in the genetic data, resulting in a likelihood surface with clear peaks for all parameters around the maximum likelihood values. The bottleneck signal
The likelihood surface is inferred from (a) 1,240 points in two iterative steps; and (b) 3,720 points in three iterative steps as described in A.3. The likelihood surface is shown only for parameter combinations that fell within the envelope of parameter points for which likelihoods were estimated. The cross denotes the maximum.

Figure 2: Examples of two-dimensional profile likelihood ratios for two data set generated with (a) $\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 40.0$ (case 0) and (b) $\theta = 0.4$, $D = 1.25$, $\theta_{anc} = 2.0$ (case 10).
is highly significant and is clearly seen in the \((\theta, \theta_{\text{anc}})\) plot on Fig. 2a, as the maximum likelihood peak is above the 1:1 diagonal. The second example is a more difficult situation, where the population has undergone a much weaker contraction \((\theta = 0.4, D = 1.25\) and \(\theta_{\text{anc}} = 2.0\), case [10]) that does not leave a clear signal in the genetic data. In such a situation, there is not much information on any of the three parameters, resulting in much flatter funnel- or cross-shaped two-dimensional likelihood surfaces. A bottleneck signal is visible on the cross-shaped \((\theta, \theta_{\text{anc}})\) plot on Fig. 2b, but is not significant.

3.2. Implementation and efficiency of IS on time-inhomogeneous models

Simulation tests show that our implementation of de Iorio and Griffiths’ IS algorithm for a model of a single population with past changes in population size and stepwise mutations is very efficient under most demographic situations tested here. Similar results are obtained for two different approximations of the likelihood (see section A.2 in the supplementary materials) First, computation times are reasonably short: for a single data set with hundred gene copies and ten loci, analyses are done within few hours to three days on a single processor, even for the longer analysis with four parameters under the GSM. Second, likelihood ratio test (LRT) p-value distributions generally indicate good CI coverage properties (see Section 5.2). Cumulative distributions of the LRT-Pvalues for all scenarios, shown in section C in the supplementary materials, are most of the time close to the 1:1 diagonal as show in Fig. 3a for our baseline scenario.
Figure 3: Cumulative distributions of P-values of Likelihood ratio tests for (a) the baseline scenario, case [0], with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 40.0$ (b) a very weak contraction scenario, case [10], with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 2.0$; and (c) a recent contraction scenario, case [3], with $\theta = 0.4$, $D = 0.125$ and $\theta_{anc} = 40.0$. Mean relative bias and relative root mean square error (RRMSE) are reported as well as the bottleneck detection rate (DR). KS indicate the Pvalue of the Kolmogorov–Smirnov test for departure of LRT-Pvalues distributions from uniformity.
Exceptions to those global trends are of two types: (1) for scenarios in which there is not much information on one or more parameters, such as the example of a weak contraction described in the previous section, likelihood surfaces are flat on the corresponding axes (Fig. 2b). Such scenarios with very few information on one or more parameters are discussed in the next section. In such situations, asymptotic LRT-Pvalue properties were not always reached (e.g. Fig. 3b) because of the small number of loci (i.e. 10) considered. Analysing more loci should improve CI coverage properties in those situations; (2) The more recent and the stronger contractions are, the less efficient are the IS proposals, because they are computed under equilibrium assumptions as detailed in Section 2.2 and A.1. Contrarily to the first situation, likelihood surfaces are then too much peaked, and maximum likelihoods are located in the wrong parameter region. The main defect we observed is thus a positive bias minimizing the contraction strength and bad CI coverage properties for $\theta$, when the number of explored ancestral histories is too small (results not shown). Consideration of 2,000 ancestral histories per parameter point (as for most simulations in this study, see Section A.1), ensures good CI coverage properties, except for some extreme situations. For a very recent and strong past contraction ($\theta = 0.4$, $D = 0.25$ and $\theta_{anc} = 400.0$), increasing the number of ancestral histories sampled for each point up to 200,000 only decreases relative bias and RRMSE on $\theta$ but does not provide satisfactory CI coverage properties (Fig. S56). Such results have however only been observed in those few extreme situations with $\theta/\theta_{anc} \leq 0.001$ and $D \leq 0.25$. Fig. 3c illustrates a more realistic situation of a very recent but not too strong population size contraction where the two defects described above are cumulated (case [3], with $\theta = 0.4$, $D = 0.125$ and $\theta_{anc} = 40.0$).

LRT-Pvalue cumulative distributions for all parameters also more often departs from the 1:1 diagonal when the mutation model moves away from a strict stepwise model and when a low number of loci (i.e. 10 or 25) is used for inference (e.g. for a GSM with $p = 0.74$ and for a K-allele model (KAM), Table 1, cases [E], [G] and [H]). In those situations, LRT-Pvalue distributions (Figs. S8, S10 and S11) often imply slightly too narrow CI, especially for parameters for which there is not much information (e.g. $\theta_{anc}$, and $D$). Considering a larger number of loci (i.e. 50) restores good CI coverage properties for the KAM but not for the GSM with $p = 0.74$ (cf. perfect LRT-Pvalue distributions for case [F] but not
for [I], Table 1, Figs. S9 and S12). This suggests that the above incorrect LRT-Pvalue distributions are partly due to the small amount of information carried by a low number of loci but also due to slight mis-specifications of the mutation model (i.e. the number of possible allelic states in the GSM, see Section A.2).
Table 1: Effects of the number of loci and mutation processes on the performance of estimations for our baseline simulation with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 40.0$ under a SMM, a GSM with $p = 0.22$ and $p = 0.74$, and a KAM, respectively.

$n_\ell$: number of loci; BDR: Bottleneck detection rate. FEDR: False expansion detection rate.

<table>
<thead>
<tr>
<th>case</th>
<th>$n_\ell$</th>
<th>$\theta$</th>
<th>$D$</th>
<th>$\theta_{anc}$</th>
<th>BDR (FEDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMM</td>
<td>[0]</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>[A]</td>
<td>25</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>[B]</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>GSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[C]</td>
<td>10</td>
<td>0.26</td>
<td>0.91</td>
<td>0.16</td>
<td>0.16</td>
</tr>
<tr>
<td>[D]</td>
<td>50</td>
<td>0.17</td>
<td>0.47</td>
<td>0.12</td>
<td>0.059</td>
</tr>
<tr>
<td>GSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[E]</td>
<td>10</td>
<td>0.016</td>
<td>0.14</td>
<td>0.0.094</td>
<td>0.137</td>
</tr>
<tr>
<td>[F]</td>
<td>50</td>
<td>0.045</td>
<td>0.081</td>
<td>3.75 $\cdot 10^{-5}$</td>
<td>0.34</td>
</tr>
<tr>
<td>KAM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>[G]</td>
<td>10</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-0.070</td>
</tr>
<tr>
<td>[H]</td>
<td>25</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-0.027</td>
</tr>
<tr>
<td>[I]</td>
<td>50</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>-0.084</td>
</tr>
</tbody>
</table>
3.3. Power and precision under ideal conditions

Results for the power of the bottleneck detection test and for the precision of the estimates under ideal conditions (i.e., same model used for simulations and analyses) with 10 loci are presented in Figs. 4 and 5.

Bottleneck detection rates (BDRs) are highest when contractions are not too recent, nor too old or too weak: a bottleneck is detected at a 5% level in more than 95% of the data sets when the contraction occurred more than 25 but less than 1400 generations ago \((0.0625 < D < 3.5, \text{Fig. 5})\), and when the ancestral population size is at least 20 times the actual size. Detection rates are then decreasing for more recent, older or weaker contractions, but stay high \((>50\%)\) in many of those situations. In this first simulation set, only extremely weak \((\theta_{\text{anc}} = 5\theta, \text{case [10], Fig. 4})\) or extremely old \((D = 7.5, \text{case [9], Fig. 5})\) contractions show bottleneck detection rates below 50%.

Precision of parameter inference is highly dependent on the scenario considered. First, global precision on all parameters increases with the strength of the bottleneck. Reasonable precision, e.g. a relative bias between -20% and 100% and RRMSE below 100%, is only obtained when the ancestral population size is greater than 20 times the actual population size (Fig. 4). However, for weaker contractions, inference of the order of magnitude for some but not all parameters can often be obtained. Second, precision of the inference of each parameter is strongly dependent on the timing of the population size change and this is well represented on Fig. 5. Parameter \(\theta\) is inferred with good precision when the contraction is not too recent, e.g. older than 200 generations in our simulation \((D > 0.5)\). For more recent contractions, relative biases are at least 130% and RRMSE larger than 300%. On the other hand, \(\theta_{\text{anc}}\) is well estimated for recent and intermediate contractions. For old contraction, e.g. older than 1,000 generations \((D > 2.5)\), relative bias and RRMSE are often greater than 100%. Inference of \(D\) shows an intermediate pattern, with more precise inferences for intermediate timings. Bias and RRMSE on \(D\) first decrease with time for contractions that occurred from 10 to 500 generations ago \((0.025 < D < 1.25)\), and then increase with time for older contractions.

Our baseline scenario, case [0] with \(D = 1.25\), thus seems to be the most favorable
Figure 4: Effect of the strength of the population size contraction on the inference of each parameter of the model (case [0] and [10] to [16]).
Relative bias are indicated by the large bars, and RRMSE by the thin lines. Stars indicate low P-values of the Kolmogorov–Smirnov test on the distribution of LRT-P-values (i.e. < 0.05). BDR: Bottleneck detection rate. FEDR: False expansion detection rate.
situation, for which inference of all parameters is relatively good given the small number of loci considered (10 loci, Figs. 3a, 4, and 5). Relative biases are only about a few percent, but RRMSE vary from 20 to 60% indicating different precision levels for the different parameters. \( D \) is the parameter inferred with the most precision, followed by \( \theta_{\text{anc}} \) and \( \theta \), respectively. Those high RMSE values are expectedly reduced when considering a larger number of loci, and reach 10 to 22% for all parameters when 50 loci are used (Table 1).

Figure 5: Effect of the timing of the population size contraction on the inference of each parameter of the model (case [0] to [9]). See Fig. 4 for details.

A few simulations have been analyzed by inferring the parameter \( D' = T \mu \) instead of
\( D = T/2N \). For those simulations, we considered \( \theta = 0.4, \theta_{\text{anc}} = 40.0 \) as in the baseline situation and four different timings (\( D = \{0.0125; 1.25; 3.5; 5.0\} \), case [17] to [20]). Our results show that scaling time by the mutation rate globally decreases the precision of the estimation of the time parameter, and does not have much effect on the other parameters \( \theta \) and \( \theta_{\text{anc}} \) (Table 2). Relative bias and RRMSE are always higher, and sometimes much higher, on \( D' \) than on \( D \). No effect of such scaling is detected on BDRs nor on the false expansion detection rate (FEDRs, results not shown).

Table 2: Effects of scaling the time by the mutation rate instead of population size for different timings, \( \theta = 0.4 \) and \( \theta_{\text{anc}} = 40.0 \). Computations are done considering only data sets with a significant bottleneck detection. No effect of such scaling is detected on BDRs nor on FEDRs.

<table>
<thead>
<tr>
<th>true ( D ) or ( D' )</th>
<th>case</th>
<th>scaling</th>
<th>( \theta )</th>
<th>( D ) or ( D' )</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.125 (0.05)</td>
<td>[3]</td>
<td>( D = T/2N )</td>
<td>2.6</td>
<td>5.7</td>
<td>( &lt; 10^{-12} )</td>
<td>0.30</td>
<td>0.65</td>
<td>2.3</td>
<td>( 10^{-4} )</td>
<td>0.040</td>
<td>0.24</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[17]</td>
<td>( D' = T/\mu )</td>
<td>2.3</td>
<td>4.5</td>
<td>( 1.4 \cdot 10^{-9} )</td>
<td>0.14</td>
<td>0.82</td>
<td>0.127</td>
<td>( -0.0026 )</td>
<td>0.46</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.25 (0.5)</td>
<td>[0]</td>
<td>( D = T/2N )</td>
<td>0.035</td>
<td>0.56</td>
<td>0.056</td>
<td>0.062</td>
<td>0.27</td>
<td>0.068</td>
<td>0.046</td>
<td>0.47</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[18]</td>
<td>( D' = T/\mu )</td>
<td>0.053</td>
<td>0.54</td>
<td>0.056</td>
<td>0.14</td>
<td>0.82</td>
<td>0.127</td>
<td>( -0.0026 )</td>
<td>0.46</td>
<td>0.857</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5 (1.4)</td>
<td>[7]</td>
<td>( D = T/2N )</td>
<td>-0.026</td>
<td>0.38</td>
<td>0.82</td>
<td>0.0038</td>
<td>0.50</td>
<td>0.51</td>
<td>0.32</td>
<td>1.7</td>
<td>0.098</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[19]</td>
<td>( D' = T/\mu )</td>
<td>-0.013</td>
<td>0.37</td>
<td>0.91</td>
<td>0.020</td>
<td>0.71</td>
<td>0.12</td>
<td>0.389</td>
<td>2.11</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.0 (2.0)</td>
<td>[8]</td>
<td>( D = T/2N )</td>
<td>-0.107</td>
<td>0.36</td>
<td>0.33</td>
<td>-0.11</td>
<td>0.42</td>
<td>0.58</td>
<td>0.46</td>
<td>2.4</td>
<td>0.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[20]</td>
<td>( D' = T/\mu )</td>
<td>-0.088</td>
<td>0.31</td>
<td>0.50</td>
<td>-0.16</td>
<td>0.52</td>
<td>0.68</td>
<td>0.49</td>
<td>2.5</td>
<td>0.60</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.4. Effect of mutational processes

To test the robustness to mutational processes, we first analyzed under a strict stepwise mutation model (SMM) samples simulated under a stable population model with \( \theta = 2.0 \) and a generalized stepwise mutation model (GSM) with \( p = 0.22 \) and 0.74 for the 10 loci considered (case [21] and [22]). For \( p = 0.22 \), 67% of the data sets show significant signals of false bottleneck. This FBDR increases up to 100% for \( p = 0.74 \). Among all simulations analyzed for this two situations, a false expansion is detected in a single data set, out of 200, with \( p = 0.22 \). The same simulations analyzed under a GSM show detection of false bottlenecks in 6 and 5% of the data sets, as well as detection of false expansion in 7.5 and 6% of the data sets, for \( p = 0.22 \) and 0.74 respectively (case [23] and [24]).
for the KAM; with 50 loci: case [D], [F] for the GSM and [I] for the KAM, Table 1).

Compared to analyses under a SMM, BDRs slightly decrease when \( p \) increases but still remain very high (e.g. \( \geq 95\% \) with 10 loci) for \( p \leq 0.74 \). On the other hand, precision of the estimations strongly differs between different parameters. Inference of \( p \) globally shows large relative bias and RRMSE for \( p = 0.22 \) but is very precise for \( p = 0.74 \). For \( \theta \), using different mutation models does not change much the precision of the estimations. For \( D \) and \( \theta_{\text{anc}} \), the mutation model has much stronger effects, showing less precise estimations for increasing \( p \) values, as well as more departure from the diagonal of the LRT-Pvalue distributions. However, increasing the number of loci from 10 to 50 restore good precision for the estimation of all parameters, except for the KAM, as well as good LRT-Pvalue distributions.

### 3.5. Effect of population structure

Table 3: Effects of isolation by distance on the detection of false contraction and expansion signals in constant-size populations. Sample scales correspond to the area (expressed as the number of lattice nodes) from which a spatially homogeneous sample is taken. \( \sigma^2 \) is the mean squared parent-offspring dispersal distance and inversely corresponds the strength of isolation by distance. See Section 5.1 for details.

<table>
<thead>
<tr>
<th>IBD strength</th>
<th>case</th>
<th>sampling scale</th>
<th>FBDR / FEDR</th>
<th>FBDR / FEDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^2 = 1 )</td>
<td>[25]</td>
<td>(10 \times 5)</td>
<td>0.67 / 0.0</td>
<td>0.16 / 0.005</td>
</tr>
<tr>
<td>( \sigma^2 = 4 )</td>
<td>[26]</td>
<td>(28 \times 28)</td>
<td>0.54 / 0.0</td>
<td>0.095 / 0.010</td>
</tr>
<tr>
<td>( \sigma^2 = 10 )</td>
<td>[27]</td>
<td></td>
<td>0.49 / 0.0</td>
<td>0.080 / 0.010</td>
</tr>
<tr>
<td>( \sigma^2 = 20 )</td>
<td>[28]</td>
<td></td>
<td>0.41 / 0.005</td>
<td>0.090 / 0.020</td>
</tr>
<tr>
<td>( \sigma^2 = 100 )</td>
<td>[29]</td>
<td></td>
<td>0.145 / 0.005</td>
<td>0.11 / 0.040</td>
</tr>
</tbody>
</table>

We first considered the presence of a local population structure by analysing samples generated under stable continuous populations with various levels of dispersal and different spatial scale of sampling (Fig. 6). Our results show that isolation by distance (IBD) structure induce high FBDRs, strongly depending on the strength of IBD as well as the spatial scale of sampling (case [25] to [29], Table 3). The stronger the IBD structure is, the higher FBDR is, varying from 15% for weak IBD with \( \sigma^2 = 100 \) to almost 70% for...
strong IBD with $\sigma^2 = 1$, for a small sampling scale. Considering larger sampling scales by sampling on the whole population area strongly decreases FBDRs to values lower or equal to 16% for all levels of IBD, but also induces false expansion detection in 4%, at most, of the data sets.
Table 4: Effects of IBD structure on the detection and characterization of a past contraction. Samples are simulated from a single continuous population under isolation by distance that has undergone a past contraction with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 40.0$. See Section 5.1 for details.

<table>
<thead>
<tr>
<th>IBD level $\sigma^2$</th>
<th>Sample scale</th>
<th>$p$</th>
<th>$\theta$</th>
<th>$D$</th>
<th>$\theta_{anc}$</th>
<th>BDR (FEDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2 = 1$</td>
<td>[30] small</td>
<td>0.71</td>
<td>1.2</td>
<td>6.6 $\cdot 10^{-9}$</td>
<td>-0.30</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>[31] large</td>
<td>0.20</td>
<td>0.88</td>
<td>0.14</td>
<td>-0.0577</td>
<td>0.46</td>
</tr>
<tr>
<td>$\sigma^2 = 4$</td>
<td>[32] small</td>
<td>0.50</td>
<td>1.1</td>
<td>5.1 $\cdot 10^{-9}$</td>
<td>-0.29</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>[33] large</td>
<td>0.22</td>
<td>0.89</td>
<td>0.74</td>
<td>-0.12</td>
<td>0.49</td>
</tr>
<tr>
<td>$\sigma^2 = 10$</td>
<td>[34] small</td>
<td>0.41</td>
<td>1.0</td>
<td>1.4 $\cdot 10^{-5}$</td>
<td>-0.19</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>[35] large</td>
<td>0.23</td>
<td>0.89</td>
<td>0.12</td>
<td>-0.11</td>
<td>0.44</td>
</tr>
<tr>
<td>$\sigma^2 = 100$</td>
<td>[36] small</td>
<td>0.35</td>
<td>0.96</td>
<td>1.6 $\cdot 10^{-4}$</td>
<td>-0.094</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>[37] large</td>
<td>0.19</td>
<td>0.86</td>
<td>0.40</td>
<td>-0.017</td>
<td>0.48</td>
</tr>
</tbody>
</table>
Using a second set of simulations under IBD with past reductions in population size, we mimic a reduction in habitat area for organisms with limited dispersal (Fig. 6). We show in Table 4 that the presence of IBD slightly decreases BDRs, e.g. from 99.5% down to 90% for very strong IBD. Strong IBD associated with small scale sampling also induces negative relative bias on $\theta$, large positive biases on $p$, $D$ and $\theta_{anc}$, as well as bad CI coverage properties as shown by KS values (Table 4) and LRT-Pvalue distributions (Figs. S42 and S44). Weaker IBD structure shows similar but weaker effects (Figs. S46 and S48). Increasing sample scale increases BDRs for situations under very strong IBD only, but strongly decreases biases and RMSE on all parameters except $\theta_{anc}$. For all situations and all parameters, considering a large sample scale allows better CI coverage properties (Table 4 and Figs. S43, S45, S47 and S49).

Table 5: Effects of an island population structure on the detection of false contraction or expansion signals in constant-size populations. Sampling scales correspond to the number of sampled demes. See Section 5.1 for details. FBDR and FEDR are respectively false bottleneck or false expansion detection rates.

<table>
<thead>
<tr>
<th>Island model settings</th>
<th>case</th>
<th>sampling scale</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>small</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 island</td>
</tr>
<tr>
<td>$\theta = 4$</td>
<td></td>
<td>FBDR / FEDR</td>
</tr>
<tr>
<td>$M = 0.01$</td>
<td>[38]</td>
<td>0.11 / 0.025</td>
</tr>
<tr>
<td>$M = 0.1$</td>
<td>[39]</td>
<td>0.32 / 0.02</td>
</tr>
<tr>
<td>$M = 1.0$</td>
<td>[40]</td>
<td>0.21 / 0.0</td>
</tr>
<tr>
<td>$M = 10.0$</td>
<td>[41]</td>
<td>0.52 / 0.0</td>
</tr>
<tr>
<td>$M = 30.0$</td>
<td>[42]</td>
<td>0.38 / 0.0</td>
</tr>
<tr>
<td>$M = 100.0$</td>
<td>[43]</td>
<td>0.19 / 0.0</td>
</tr>
<tr>
<td>$\theta = 20$</td>
<td>$M = 1.0$</td>
<td>[44]</td>
</tr>
</tbody>
</table>

We finally tested the influence of an island population structure with varying levels of migration and population sizes (Table 5 and 6). Our results first show that sampling a single island from a stable structured population induces high FBDRs, from 11 to 52% depending on the level of population structure. With such local sampling scheme, the relationship between FBDRs and the level of population structure is complex. Increasing sampling scale by sampling three to ten populations instead of a single one has two major antagonistic
effects: it strongly increases FBDRs up to 100% for highly structured situations (i.e. \( M \leq 1.0 \)); but it also reduces FBDRs for less structured populations, down to values around 10% for the larger sampling scale. Note that a few false expansions were also detected among all those simulations but always in less than 5% of the data sets. Finally, we can also note that, in our simulations, increasing the total diversity \( \theta \) strongly increases the effect of population structure except when all demes are sampled but more simulations would be needed to conclude on the general effect of genetic diversity.

Our last set of simulations under an island model with past reduction of population sizes shows an extremely strong effect of the sampling scale and the level of population structure (Table 6). Compared to unstructured situations with BDR=99% (case [G]), sampling a single deme strongly reduces BDRs to values from 88% for weak population structure with \( M = 100.0 \), down to 0.5% for highly structured populations with \( M = 0.01 \). Intermediate structure with \( M = 1.0 \) also lead to small BDR of 7% in our simulations. With such a small sampling scale, parameter estimation is clearly inaccurate when population structure is not very weak (e.g. \( M \ll 100.0 \)), showing strong bias, large RMSE, and bad coverage properties of CIs (Table 6, and Figs. S50 to S53). This is observed whatever expected values, i.e. local for deme values or global for the whole population values, are considered (results not shown). The effect however decreases with higher levels of migration, and parameter inference is relatively accurate with \( M = 100.0 \) for all parameters except \( p \), and shows reasonable CI coverage properties (Figs. S54 and S55). Increasing sampling scale clearly increases BDRs but, contrarily to the results obtained for IBD, it does not improve parameter inferences nor CIs coverage properties. Sampling a large scale seems to allow better estimation of \( \theta_{anc} \), but the effect on all other parameters is highly dependent on the demographic scenario considered, and no clear conclusion can be drawn from our simulations.
Table 6: Effects of an island population structure on the detection and characterization of a past contraction. Samples are simulated from an 10-island model in which each sub-population has undergone a past contraction, with $\theta = 2dN_d \mu = 0.4$, $D \equiv T/(2dN_d) = 1.25$ and $\theta_{anc} = 2dN_{d,anc} \mu = 40.0$, and varying scaled migration rate $M = 2N_d m$. See Section 5.1 for details. Mean relative bias and Relative Root Mean Square Error (RRMSE) are reported as well as the bottleneck detection rate (BDR) and the false expansion detection rate (FEDR). KS indicate the $P$ value of the Kolmogorov–Smirnov test for departure of LRT-Pvalues distributions from uniformity.

<table>
<thead>
<tr>
<th>Gene flow level</th>
<th>case</th>
<th>Sampling scale</th>
<th>$P$</th>
<th>$\theta$</th>
<th>$D$</th>
<th>$\theta_{anc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M = 0.01$</td>
<td>[45] small</td>
<td>-0.081</td>
<td>1.0</td>
<td>0.026</td>
<td>-0.61</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>[46] very large</td>
<td>2.2</td>
<td>2.3</td>
<td>2</td>
<td>-0.77</td>
<td>0.81</td>
</tr>
<tr>
<td>$M = 1.0$</td>
<td>[47] small</td>
<td>1.6</td>
<td>1.9</td>
<td>&lt; 10^{-12}</td>
<td>-0.32</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>[48] very large</td>
<td>-0.0027</td>
<td>0.61</td>
<td>4.2 · 10^{-8}</td>
<td>-0.72</td>
<td>0.80</td>
</tr>
<tr>
<td>$M = 100$</td>
<td>[49] small</td>
<td>0.69</td>
<td>1.2</td>
<td>4.2 · 10^{-6}</td>
<td>-0.070</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>[50] very large</td>
<td>0.014</td>
<td>0.83</td>
<td>0.0085</td>
<td>0.13</td>
<td>0.55</td>
</tr>
</tbody>
</table>

KS indicate the $P$ value of the Kolmogorov–Smirnov test for departure of LRT-Pvalues distributions from uniformity.
4. Discussion

In this study, we adapted de Iorio and Griffiths’ IS algorithm to consider a single population model with varying size and different mutation models. We investigated its performance in detecting past contractions of population size as well as estimating the model parameters. We focussed on the effect of the timing and amplitude of the past population size contraction, and on the robustness of inferences to mis-specifications of the mutational and population structure models. Our results allow us to illustrate both the strengths and imperfections of the method.

4.1. Performances under ideal conditions

First, over all simulations considered in this study, LRTs for CI coverage indicate that our implementation is correct and produces accurate estimates of the likelihood surface with reasonable computation times, except in a few situations with extremely strong demographic disequilibrium (i.e. for recent, e.g. $D \leq 0.25$, and strong bottlenecks, e.g. $N_{anc}/N \geq 1,000$). For the later situations, much longer runs are needed to obtain good CI coverage. This shows that the efficiency of de Iorio and Griffiths’ IS algorithm, based on time-homogeneous demographic assumptions, strongly depends on the extent of the demographic disequilibrium considered, which can be roughly quantified by the ratio of the amplitude of the population size change divided by its duration. Our results also show that inference based on time-homogeneous IS algorithms is practically intractable for the most extreme situations.

Second, our simulations show very good performances in terms of detection of past decreases in population size. BDRs are larger than 95% for most demographic situations. Even very recent (e.g. $T = 10$ generations, $D = 0.025$), relatively ancient (e.g. $T = 2,000$ generations, $D = 5.0$) or relatively weak contractions (e.g. population size ratio of 10) are detected in more than 50% of the data sets. Third, our results suggest that using only 10 microsatellite markers allows detecting past contraction with a high power, but more markers are required for precise inferences of scaled population sizes and timing under a wide range of demographic situations. However, precision of the inference of the different parameters strongly depends on the scenario considered (see also Girod et al., 2011). This is not surprising because the performance of the method to infer past demography
strongly depends on the genetic information available in the data and this information
strongly varies as a function of the timing of the past contraction. This can easily be
understood and predicted from the timing of events in the ancestry of a sample. Recent
contractions result in more precise inference for the ancestral population size than for the
actual population size, because much of the coalescent and mutation events occur in the
ancestral population. The opposite is true for old contractions. Precise inference of current
and ancestral population sizes is thus only expected for past contractions that did neither
occurred too recently nor too far in the past because in such scenarios coalescent and mu-
tation events are more homogeneously distributed over all demographic phases. Finally
and for the same reason, inference of contraction time is expected to be more precise for
intermediate timings. This is exactly what is observed in Figs. 4 and 5. One important
result of our study is that high BDRs as well as good inference precision for both the time
and the actual population size parameters is still expected for relatively ancient contraction
(e.g. $1.25 \leq D \leq 5.0$).

Finally, many recent programs that make demographic inferences from genetic data,
such as MIGRATE (Beerli and Felsenstein, 2001), IM (Hey and Nielsen, 2004, 2007; Hey,
2010), or LAMARC (Kuhner, 2006) software packages, do not use the classical coalescent
parameter scaling by population size (i.e. $4Nm$ and $T/2N$) but rather use scaling by the
mutation rate (i.e. $m/\mu$ and $T\mu$), or propose both options, as MIGRAINE does. Our simu-
lations show that there is not much interest to scale time by mutation rate for inferences
of past bottlenecks. For the demographic scenarios considered here, such scaling always
reduce inference precision for the time parameter. Beside those scaling issues, independent
information about mutation rates of the makers can be incorporated as prior information
in the analyses to allow inference of canonical parameters (i.e. $N$, $T$ and $N_{\text{anc}}$) instead of
scaled ones (e.g. as done in MSVAR). This is an attractive possibility for practical inferences,
however, it has been shown in Girod et al. (2011) and Faurby and Pertoldi (2012) that
such parametrization allows precise inference of canonical parameters only if precise prior
information on mutation rate is used. This is so because single-locus population genetics
models in general (and Kingman’s coalescent model in particular) depend upon scaled, not
canonical, parameters.
4.2. Comparison with previous methods

In the past decade, the use of likelihood-based methods to analyze genetic data under a single population model with past variation in population size emerged with the release of the MsVar software (Beaumont, 1999; Storz and Beaumont, 2002). As expected, this coalescent-based MCMC method has been shown to be much more powerful than using summary statistics in detecting past bottlenecks or expansions (Girod et al., 2011; Peery et al., 2012). Moreover, model-based approaches can also infer model parameters, such as current and past population sizes, and the timing of the demographic change. In this study, we compared performances in terms of bottleneck detection rates, parameter inference precision and computation times of MsVar and our IS method. Our simulations globally show similar behavior of the two methods, with slight but clear advantages for Migraine in terms of power of bottleneck detection, parameter estimation and computation times. For example, both methods are very performant for intermediate bottleneck strength and timing. On the contrary, they both are inefficient when population size contraction is too strong and too recent. The MCMC algorithm of MsVar show strong convergence issues for very recent and strong bottlenecks (see Fig.1 in Girod et al., 2011) and give biased point estimates as well as bad CI for $\theta_{anc}$ (Fig. S3). For the same demographic scenarios, IS algorithms implemented in Migraine are not efficient and, even with large computation times, Migraine show high relative biases and RRMSE, as well as bad coverage properties of CI for $\theta$ as already shown in Section 3.2 and discussed above. For both methods, computation times thus greatly increase with the strength of the bottleneck, and accurate parameter inference considering very recent and strong bottlenecks may be difficult to achieve. However, Migraine appears (1) more adapted to the analyses of microsatellite markers because of the implementation of the GSM model; (2) slightly more powerful than MsVar as our simulations show higher bottleneck detection rates and less false expansion detections; and (3) faster than MsVar as computation times were always higher for MsVar than for Migraine for equivalent demographic scenarios (e.g. two to ten times faster). Finally, a certain advantage of Migraine over MsVar is that it can easily use parallel computation, thereby decreasing computation times by the the number of available cores.
4.3. Robustness to mutational processes

Although many models have been developed to describe microsatellite mutation processes (Bhargava and Fuentes, 2010), most programs that analyze microsatellite data use the SMM (e.g. IM, MIGRATE, LAMARK, see references above, but see DIYABC, Cornuet et al., 2008, and BEAST, Drummond et al., 2012). However, it has been recognized that violations of the SMM assumptions might induce severe bias in the inference of demographic history (Gonser et al., 2000). Indeed, mutations of more than one step of the GSM can produce gaps in allele length distribution, which are typically often observed after a population decline under a SMM (Garza and Williamson, 2001). Peery et al. (2012) recently showed that identification of past bottlenecks using the summary statistic-based Bottleneck (Cornuet and Luikart, 1996) and M-Ratio softwares (Garza and Williamson, 2001), are highly biased by deviations from the mutation models implemented in those softwares, often leading to significant bottleneck detection in samples simulated under a stable population model. Faurby and Pertoldi (2012) also showed that estimation of present and past population sizes with MsVAR are unreliable when realistic deviations from the SMM occurs. In the previous study of Girod et al. (2011), we also showed that MsVAR was moderately robust to deviations from the SMM, which lead to false bottleneck detections in sample simulated from stable populations. However, this conclusion was presumably over-optimistic due to the small number of data sets analyzed. In the present study, we clearly show a strong impact of violations of the SMM assumptions: even small deviations from the SMM induce large FBDRs in samples simulated under a stable demography. We adapted our algorithm by implementing a GSM in Migraine to allow inference of past population size variations under this more complex and more realistic mutational model. Our simulations under a GSM show BDRs similar to the one observed under the SMM, and also show that parameter inference precision is only slightly affected by the additional parameter. Computation times are of course higher than for the SMM, but are still reasonable, as all data sets with 100 genes genotyped at 50 loci or less can be analyzed in a few days on a desktop computer with 4 cores. It is clear however that the GSM is not a perfect description of microsatellite mutation processes (Bhargava and Fuentes, 2010). Many other factors such as single nucleotide insertions/deletions, asymmetric mutations, variation of the mutation rate with the length of the alleles and/or constraints on allele
sizes may often occur (Sun et al., 2012) and potential additional biases in the inference of past population size changes due to these factors remain to be tested. However, the main cause of the confounding effects between mutation processes and past changes in population sizes is likely to be the presence of gaps in the sample allelic distributions, and factors others than multistep mutations should have less effects than those described above.

4.4. Robustness to population structure

In addition to the strong effect of mutational processes, we also found that inferences of past population size changes can be drastically affected by population structure. First, at small spatial scales, isolation by distance often occurs within populations due to spatially limited dispersal (see Guillot et al., 2009 for a review). Our simulations show that ignoring such local population structure induce large FBDRs when individuals are sampled at a small spatial scale from stable populations, even for relatively weak IBD. However, sampling individuals at a larger scale, i.e. over the whole population area, efficiently reduces FBDRs. Parameter estimation, and to a lesser extent BDRs, obtained from samples coming from a population that effectively went through a bottleneck are also affected by IBD, and again sampling individuals at a large geographical scale efficiently reduce the impact of IBD. Parameter inference thus appears robust unless IBD is very strong and sampling scale small.

Second, at larger spatial scales, population structure also arises due to limited gene flow within a set of discrete demes as described by the island model (Wright, 1951). Such island population structure has stronger and more complex effects than IBD within populations. Our simulations show that samples coming from a single deme of stable island-structured populations show large FBDRs unless gene flow is extremely limited. Considering larger sampling scales by sampling individuals from all the demes of the total structured population reduces FBDRs but only for situations with important levels of gene flow (i.e. $M \geq 10.0$). Contrarily to IBD situations, enlarging sampling scale when gene flow is more limited, i.e. $M \leq 1.0$, always increases FBDRs. Our simulations finally show that, when a bottleneck did occur in the past, ignoring island population structure also often strongly decreases BDRs and greatly biases parameter estimation. Moreover, both bottleneck detection and parameter estimation are sensitive to sampling scale. Unless gene flow is very high
between demes ($M \geq 100.0$), small scale samples show low BDRs below 10% and accurate
BDRs are only obtained using large sampling scales. Parameter inference appears highly
biased for all levels of gene flow considered in this study. As for BDRs, best precision is
also obtained when gene flow is high and sampling scale is large. Nevertheless, for all other
situations, relative biases and RRMSEs are high suggesting that in most situations, limited
gene flow between geographically distinct demes will always lead to erroneous inferences
of past and present population sizes, and of the timing of the demographic change.

Such confounding effects of population structure and past changes in population sizes
has already been observed. First, the effect of small scale IBD population structure on
BDRs obtained with the **BOTTLENECK** and **M-Ratio** softwares has been tested by simulations
in Leblois, Estoup and Streiff (2006). Our results are globally in agreement with this
previous study, except that they found large FEDRs when using **BOTTLENECK** on IBD
samples and that considering large scale samples makes FEDRs even larger. Such results
showing that fine scale population structure induces false expansion signals has also been
previously stressed by Ptak and Przeworski (2002) in the context of sequence data analysis
based on the Tajima’s $D$ statistics. Our simulations on the contrary show non-null but
small FEDR in the presence of small scale IBD structure.

Second, the effect of island population structure on past population size inference was
first highlighted by simulation in Nielsen and Beaumont (2009). More recently, Peter,
Wegmann and Excoffier (2010), Chikhi et al. (2010) and Heller, Chikhi and Siegismund
(2013) also showed that analyzing samples drawn from a single deme of an island model
with low to intermediate migration rates (i.e. $Nm < 5$) leads to false signals of bottleneck.
Such erroneous imputations can be understood by considering the genealogical processes
in an island model and in a single population with varying size. In a subdivided popula-
tion with relatively small deme sizes and small migration rates, the genealogy of a sample
taken from a single deme will show (1) many short branches for genes that rapidly coalesce
within the deme in which they were sampled (i.e. before any migration event), this corre-
respond to the “scattering phase” described in Wakeley (1999); and (2) a few much longer
branches for genes that coalesce after any emigration or immigration event from the deme
sampled, this is the “collecting phase” of Wakeley (1999). The result is a genealogy with
an excess of short terminal branches, as expected after a recent contraction in population
size. However, if only one individual is taken from different demes, and/or if deme size or migration rates are large, the genealogical process becomes closer to the one expected under a Wright-Fisher population. Similarly, when gene flow is very limited, the ancestry of a sample coming from a single deme will also be very similar to the one expected under the WF model. Thus, except for limit cases, structured and declining population scenarios may result in more or less similar genealogies, depending on deme sizes, migration rates and sampling scale. This expected influence of these three factors may strongly complicate the study of the effect of population structure on the inference of past population size. This can be noticed in the heterogeneity of the results of the different simulation studies available. All those comparisons based on different simulations of structured population show that the effect of population structure is generally complex, and will be quite difficult to predict except in a few simple cases. Those results also show that verbal argumentation based on over-simplified past genealogical processes may not always give the right prediction. Nevertheless, three main points arise from those simulation studies and can serve as guidelines for empirical studies: (1) using a large sample scale strongly limits the influence of population structure on the inference of past population size variations, as advocated by Chikhi et al. (2010), but allows correct inference only when a single individual (ideally, a single gene) is sampled per deme or when migration rates are relatively high, i.e. $M \geq 10.0$; (2) for all other demographic situations, detection of past population size changes and parameter inferences based on panmictic models may often be misleading. However, we did not here consider sampling a single individual per deme, which may more effectively decrease the bias due to population structure.

Such results finally implies that models themselves should be improved. First, model choice procedure should be developed to evaluate whether observed patterns of genetic diversity can be better explained by a model of population size change or by a model of subdivided populations. For example, Peter, Wegmann and Excoffier (2010) used an Approximate Bayesian Computation (ABC) model choice approach to distinguish between structured populations and panmictic population that undergone past changes in size. However, they show by simulation that their model choice procedure has relatively limited power to assign simulated data sets to the correct evolutionary model, even with a relatively large number of loci (e.g. 60% to 85.5% with 10 to 200 loci, respectively). An
alternative is to develop models accounting for both population structure and population size changes would probably be more realistic for most species/populations but the only available method (Hey and Nielsen, 2007; Hey, 2010) has never been tested for scenarios with both structured populations and past changes in population sizes.

4.5. Conclusion

This work shows that our new inference method seems very competitive compared to alternative methods, such as \texttt{MsVar}. However, our simulation tests also showed some important limits, which most importantly are large computation times for strong disequilibrium scenarios and a strong influence of some form of unaccounted population structure. One first major improvement would thus be to speed up the analyses. Among the different possibilities, a relatively simple improvement would be to more efficiently choose the number of explored histories for each point of the parameter space. A more attractive improvement would be to design more efficient IS algorithms for time-inhomogeneous models. However, various unsuccessful attempts suggest that it may be a difficult task (not shown). A second major improvement would be to include population structure in the demographic model for simultaneous inference of migration rates and past population size change or to develop model choice procedures.

Lastly, given the current revolution in genetic data production due to next generation sequencing technologies (NGS), it seems crucial to allow for the analysis of different types of independent markers, such as small DNA sequences without intra-locus recombination, or SNPs. Given the relatively large computation times of our method, all analyses will clearly only be tractable for a limited number of markers (e.g. \textless 10,000), but could nevertheless give very precise inferences. However, considering only independent markers is probably not the optimal approach as NGS make it possible to apply new class of methods based on the analyses of linkage disequilibrium for past demographic inferences. Such methods are based on the computation of the distribution of non-recombining haplotype block length (e.g. Meuwissen and Goddard (2007); Albrechtsen et al. (2009); Gusev et al. (2012); Palamara et al. (2012); Theunert et al. (2012) or explicitly model the spatial dependence of markers using hidden Markov models (e.g. Dutheil et al. (2009); Mailund et al. (2012)). They will probably play a major role in the future of population genetic demographic and
5. Methods

5.1. Simulation study

Table 7: Simulated demographic scenarios with a stepwise mutation model

<table>
<thead>
<tr>
<th>Case</th>
<th>$D$ ($T$)</th>
<th>$\theta$ ($N$)</th>
<th>$\theta_{anc}$ ($N_{anc}$)</th>
<th>Case</th>
<th>$D$ ($T$)</th>
<th>$\theta$ ($N$)</th>
<th>$\theta_{anc}$ ($N_{anc}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[0]</td>
<td>7.5 (3,000)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
</tr>
<tr>
<td>[1]</td>
<td>0.025 (10)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[10]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>2.0 (1,000)</td>
</tr>
<tr>
<td>[2]</td>
<td>0.0625 (25)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[11]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>4.0 (2,000)</td>
</tr>
<tr>
<td>[3]</td>
<td>0.125 (50)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[12]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>8.0 (4,000)</td>
</tr>
<tr>
<td>[4]</td>
<td>0.25 (100)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[13]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>12.0 (6,000)</td>
</tr>
<tr>
<td>[5]</td>
<td>0.5 (200)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[14]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>24.0 (16,000)</td>
</tr>
<tr>
<td>[6]</td>
<td>2.5 (1,000)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[15]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>120.0 (60,000)</td>
</tr>
<tr>
<td>[7]</td>
<td>3.5 (1,400)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td>[16]</td>
<td>1.25 (200)</td>
<td>0.4 (200)</td>
<td>400.0 (200,000)</td>
</tr>
<tr>
<td>[8]</td>
<td>5 (2,000)</td>
<td>0.4 (200)</td>
<td>40.0 (20,000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A first set of simulations aims at testing the power of the algorithm to detect bottlenecks and the accuracy of the parameters estimates when the duration ($D$) or the strength of the contraction ($\theta_{anc}$) vary. The mutation process considered is a SMM over a range of 200 alleles. These experiments are presented in Table 7. We also reanalyzed the sixty simulated data sets from Girod et al. (2011) to compare the results obtained with MsVAR and our own estimates. The latter simulated data sets are described in Table S2 and the comparison results are presented in section B in the supplementary materials.

A second set of simulations concerns robustness and accuracy related to mutation processes of microsatellites that are known to be highly complex (Ellegren, 2000, 2004; Sun et al., 2012). This second set of simulations is thus based on a generalized stepwise mutation model (GSM) with either $p = 0.22$ or $p = 0.74$, that are respectively the value commonly considered as a realistic average value in the literature (Dib et al., 1996; Ellegren, 2000; Estoup et al., 2001; Ellegren, 2004), and the largest, ever reported value (Fitzsimmons, 1998; Peery et al., 2012). We have also added data sets drawn with the $K$-allele model (KAM) to those simulations, which might be seen as a GSM with $p = 1.0$. A first set
Figure 6: Simulated data sets with population structure

Local population structure

The simulated IBD populations are composed of individuals set at the nodes of a regular lattice, whose size can vary. A past reduction in population size is thus modeled as a reduction of the habitat area keeping a constant density of individuals. Various levels of localized dispersal were simulated via truncated Pareto distributions with mean squared parent-offspring dispersal distance, say $\sigma^2$, varying in \{1; 4; 10; 100\}.

Parameters of the IBD populations

- At equilibrium: $\theta = 4.0$ with a $32 \times 31$ lattice (hence $N = 1984$ genes)
- Including an habitat contraction: $(D, \theta, \theta_{anc}) = (1.25, 0.4, 40.0)$ with lattices of sizes from $10 \times 10$ ($N = 200$) to $100 \times 100$ ($N_{anc} = 20,000$) backward in time

Simulated sampling schemes

100 genes sampled

- on a $5 \times 10$ lattice in the center of the population [small sample scale], or
- regularly on the whole area (i.e., one individual every 4 nodes) [large sample scale].

Island population structure

We considered models with $d = 10$ demes of equal size $N_d$ genes, varying in \{20; 200; 2000\}, and exchanging migrants at rate $m$ between pairs of demes, varying in \{0.000025; 0.00025; 0.0025; 0.025; 0.075; 0.25\}. The model is fully characterized by the scaled parameters $\theta = 2dN_d\mu$ and $M = 2N_d m$. When past contractions occurred, deme sizes $N_d$ decreased forward in time but migration rates $m$ are kept constant in time. Values of $M$ reported below correspond to scaled migration rates at sampling time $t = 0$.

Parameters of the island populations

- $\theta \in \{4.0, 20.0\}$ and $M \in \{0.01, 0.1, 1.0, 10.0, 30.0, 100.0\}$ without population size changes
- $\theta = 0.4$, $M \in \{0.01, 1.0, 100.0\}$ and a contraction with parameters $D = 1.25$, $\theta_{anc} = 40.0$

Simulated sampling schemes

Samples of 100 genes picked at random

- from a single deme [small sample scale], or
- from three demes [large sample scale], or
- from all demes [very large sample scale].
of analyses tests the robustness to mis-specification of the mutation process. Indeed, we have simulated under a GSM but inferred under a SMM. A second set of analyses tests the accuracy of the estimates when the inference algorithm is based on a GSM with unknown value of \( p \).

The aim of the third set of simulations is to test robustness against a population structure that is ignored by the inference algorithm. All the data sets in this last series were simulated under a GSM model with \( p = 0.22 \) and are presented in Table 6. A first group of data sets simulates local within-population structure according to an isolation-by-distance (IBD) model. It thus aims at testing the robustness of the inferences to the assumption of panmixture by considering non-random mating due to spatially localized parent-offspring dispersal. The second group of data sets simulates both within- and among-population structure at a larger spatial scale according to an island model.

For each scenario, we simulated 200 multilocus data sets. Each simulated data set is a sample of \( n_g = 100 \) genes (or haploid individuals), genotyped at \( n_\ell = 10 \) unlinked microsatellite loci, except for a few situations where we indicate that 25 or 50 instead of 10 loci are used. The mutation rate per gene per generation, say \( \mu \), is assumed to be constant for all loci, equal to \( 10^{-3} \). All simulated samples, except data sets from Girod et al. (2011) (see section B in the supplementary materials), have been produced with a new version of the \textsc{IBDSim} software (Leblois, Estoup and Rousset, 2009) that considers continuous changes of population sizes.

5.2. Validation

In all simulation experiments, the true (simulated) values of the parameters of interest are compared to the estimated values. The estimation bias and error, assessed by the relative mean bias and relative root mean square error (RRMSE), are reported, as well as the proportion of data sets for which a bottleneck or a false expansion signal is significantly detected (BDR and FEDR, respectively). Furthermore, the accuracy of the inference methodology is assessed by mean of profile likelihood ratio tests (LRTs, Cox and Hinkley, 1974; Severini, 2000). The coverage properties of the confidence intervals computed from the smoothed likelihood surface are tested via the distributions of LRT p-values, which should be asymptotically uniform. The departure from uniformity is tested.
by Kolmogorov–Smirnov tests, notably to check the validity of the implementation of the inference method and to assess the different factors that can affect likelihood surface inference.

The supplementary materials are available at the XXX website. The MIGRAINE software, with the implementation of the above described methods, can be downloaded from the web site kimura.univ-montp2.fr/~rousset/Migraine.htm.

Acknowledgements

We are grateful to L. Chikhi, J.-M. Cornuet, A. Estoup, J.-M. Marin for their constructive discussions about this work. This study was supported by the Agence Nationale de la Recherche (EMILE 09-blan-0145-01 and IM-Model@CORAL.FISH 2010-BLAN-1726-01 projects) and by the Institut National de Recherche en Agronomie (Project INRA Starting Group “IGGiPop”). Part of this work was carried out by using the resources of the Computational Biology Service Unit from the MNHN (CNRS Unité Mixte de Service 2700), the INRA MIGALE and GENOTOUL bioinformatics platforms and the computing grids of ISEM and CBGP labs.

References


Heller R, Chikhi L, Siegismund HR, 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. PloS one, 8:e62992.


A. Details on the likelihood computations and settings of the inference method

A.1. Coalescent-based IS algorithms and disequilibrium models

In this section, we give a more detailed overview of the method used to compute likelihoods of genetic data at a given locus and technical details regarding the Monte Carlo algorithm.

The likelihood at a given point of the parameter space is estimated using Stephens and Donnelly (2000) and de Iorio and Griffiths (2004a)'s importance sampling approach. An ancestral history, i.e. a coalescence tree with mutations, is defined as the set of all ancestral configurations $H = \{H_k; k = 0, -1, ..., -m\}$, corresponding to all coalescent or mutation events that occurred from $H_0$ the current sample state (i.e. the sample allelic configuration, or allelic counts) to $H_{-m}$ the allelic state of the most recent common ancestor (MRCA) of the sample. The Markov nature of the backward coalescent process implies that $p(H_k) = \sum_{H_{k-1}} p(H_k|H_{k-1})p(H_{k-1})$ and expending the recursion over possible ancestral histories of a current sample leads to $p(H_0) = E_p[p(H_0|H_{-1})...p(H_{-m+1}|p(H_{-m})].$

However, forward transition probabilities $p(H_k|H_{k-1})$ can not directly be used in a backward process and backward transition probabilities $p(H_{k-1}|H_k)$ are unknown, except in some specific simple models such as parent independent mutations (PIM) in a single stable panmictic population. Importance sampling techniques based on an approximation $\hat{p}(H_{k-1}|H_k)$ of $p(H_{k-1}|H_k)$ are thus used to derive the probability of a sample over possible histories

$$p(H_0) = E_p\left[p(H_0|H_{-1})...p(H_{-m+1}|H_{-m})\hat{p}(H_{-1}|H_0)...\hat{p}(H_{-m}|H_{-m+1})\right].$$

The likelihood of the data is then estimated as the average value of the probability of a sample configuration $H_0$ given an ancestral history $H_i$, over $n_H$ independent simulations.
backward in time, of possible ancestral histories: \( p(H_0) \approx \frac{1}{n_H^k} \sum_{i=1}^{n_H^k} p(H_0|H_i) \). The distribution of the possible ancestral histories is generated by an absorbing Markov chain with transition probabilities \( \hat{p}(H_{k-1}|H_k) \), and the likelihood is estimated by averaging the product of sequential importance weights corresponding the ratio \( \frac{p(H_{k,0}|H_{k,1}) \ldots p(H_{i,m-1}|H_{i,m})}{\hat{p}(H_{k,1}|H_{k,0}) \ldots \hat{p}(H_{i,m-1}|H_{i,m})} \) of forward and backward transition probabilities obtained for each history \( H_i \). Computation of backward transition probabilities \( \hat{p} \) relies on Stephens and Donnelly’s \( \hat{\pi} \) approximations of the unknown probability \( \pi \) that an additional gene sampled from a population is of a given allelic type conditional on a previous sample configuration (see A.4 for an example on \( \hat{\pi} \) computations).

For efficient importance sampling distributions (i.e. approximate backward transition probabilities \( \hat{p} \) close to the exact \( p \)), precise estimation of the likelihood can be obtained with very few histories explored. For example, under parent independent mutation model (e.g. a K allele model, KAM), the importance sampling scheme of Stephens and Donnelly (2000) for a single isolated population is optimal because the \( \hat{\pi} \)'s computed following Stephens and Donnelly (2000) are equal to the true \( \pi \)'s and all the ratios of \( \frac{p(H_{k,1}|H_{k,0})}{\hat{p}(H_{k,1}|H_{k,0})} \) are cancelled out. In such ideal case, consideration of a single ancestral history is thus sufficient to get the exact likelihood (de Iorio and Griffiths, 2004a,b; Stephens and Donnelly, 2000). However, departure from parent independent mutation, from panmixia and from time-homogeneity decrease the efficiency of IS proposals, and precise estimation of likelihoods then implies to explore more ancestral histories. Under a time-homogeneous model of isolation by distance, Rousset and Leblois (2007, 2012) found that 30 replicates of the absorbing Markov chain, rebuilding 30 independent possible ancestral histories, is enough to get perfect LRT-Pvalue distributions. Here, for the first time, de Iorio and Griffiths (2004a)'s IS algorithm is applied to a time-inhomogeneous demographic model using \( \hat{\pi} \) probabilities computed from a time-homogeneous demographic model as described in the main text. Our work clearly shows that IS proposal computed as described in the main text is less and less efficient for demographic scenarios with increasing disequilibrium.

A.2. Some efficient modifications of the original IS algorithm

To speed up the computations, we can stop the simulation of the genealogies during the IS algorithm, before reaching the MRCA (see Jasra, De Iorio and Chadeau-Hyam, 2011; Rousset and Leblois, 2007). Here, the algorithm is stopped when reaching demographic equilibrium (i.e., after time \( T \) when \( N(t) = N_{anc} \)) and we finalize the current IS estimates of the likelihood with the PAC-likelihood of the ancestral lineages (Li and Stephens, 2003; Cornuet and Beaumont, 2007; Rousset and Leblois, 2007, 2012). This scheme will be called PACanc. Using analytical formulas for the exact probability of the last pair of genes, computed as in Rousset (2004), also slightly decreases the computation time in the same vein as in de Iorio et al. (2005) and Rousset and Leblois (2007, 2012). The IS scheme can thus be stopped when reaching an ancestral sample of size 2 and finalized with this exact formula, hereafter called 2ID. And the combination of both PACanc and 2ID will be called PACanc2ID. A detailed comparison between the four schemes (strict IS, 2ID, PACanc and PACanc2ID) is presented is presented in Table S1 for the baseline scenarios under a SMM and under a GSM (case[0], [A] to [F] under a SMM; case[G] to [L] under a GSM with \( p = 0.22 \); and case[K] to [M] under a GSM with \( p = 0.74 \)). For the GSM, data sets are simulated under a GSM with 40 allelic states but analyzed with a GSM with 50 possible
allelic states. This slight mis-specification of the mutation model has a relatively strong influence when \( p = 0.74 \): LRT-Pvalue distributions are not close tho the 1:1 regardless of the number of loci. Such strong effect is also probably due to the consideration of large \( \theta_{anc} \) values. Apart from this mutation model effect, our results show that performances are similar in all cases for the different algorithms. All simulations with a GSM, i.e. for the tests of the effect of mutational processes and population structure, are analyzed using the PACanc2ID, unless otherwise specified.
Table S1: Effects of the algorithm used and the number of loci on the performance of estimations for our baseline simulation with $\theta = 0.4$, $D = 1.25$ and $\theta_{anc} = 40.0$ under a SMM, a GSM with $p = 0.22$ and $p = 0.74$. $\hat{L}$: type of algorithm used; $n_L$: number of loci; IS: importance sampling; 2ID: exact computation of the likelihood for the last pair of genes; PACanc: PAC-likelihood used in the ancestral stable population. BDR: Bottleneck detection rate. FEDR: False expansion detection rate.

<table>
<thead>
<tr>
<th>case / $\hat{L}$</th>
<th>$n_L$</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>rel. bias</th>
<th>RRMSE</th>
<th>KS</th>
<th>BDR (FEDR)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SMM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J IS</td>
<td>10</td>
<td>0.23</td>
<td>0.51</td>
<td>0.12</td>
<td>0.16</td>
<td>0.66</td>
<td>0.14</td>
<td>0.22</td>
<td>1.34</td>
<td>0.65</td>
<td>0.30</td>
<td>0.32</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>J 2D</td>
<td>10</td>
<td>0.24</td>
<td>0.51</td>
<td>0.37</td>
<td>0.091</td>
<td>0.48</td>
<td>0.1</td>
<td>0.10</td>
<td>1.06</td>
<td>0.028</td>
<td>0.99</td>
<td>0.003</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>L PACanc2ID</td>
<td>10</td>
<td>0.21</td>
<td>0.52</td>
<td>0.36</td>
<td>0.17</td>
<td>0.59</td>
<td>0.37</td>
<td>0.21</td>
<td>1.16</td>
<td>0.4</td>
<td>0.985</td>
<td>0.003</td>
<td>0.985</td>
<td></td>
</tr>
<tr>
<td>A IS</td>
<td>25</td>
<td>0.21</td>
<td>0.51</td>
<td>0.37</td>
<td>0.11</td>
<td>0.49</td>
<td>0.53</td>
<td>0.11</td>
<td>1.0</td>
<td>0.675</td>
<td>0.99</td>
<td>0.003</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>M PACanc2ID</td>
<td>25</td>
<td>0.17</td>
<td>0.51</td>
<td>0.37</td>
<td>0.012</td>
<td>0.48</td>
<td>0.75</td>
<td>0.085</td>
<td>0.39</td>
<td>0.082</td>
<td>1.0</td>
<td>0.003</td>
<td>1.0</td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>50</td>
<td>0.13</td>
<td>0.51</td>
<td>0.37</td>
<td>-0.13</td>
<td>0.37</td>
<td>0.11</td>
<td>0.10</td>
<td>1.06</td>
<td>0.28</td>
<td>0.99</td>
<td>0.003</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>IS 0.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>10</td>
<td>0.13</td>
<td>0.57</td>
<td>0.53</td>
<td>0.42</td>
<td>0.67</td>
<td>10^{-12}</td>
<td>2.46</td>
<td>3.4</td>
<td>10^{-12}</td>
<td>0.965</td>
<td>0.965</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS</td>
<td>25</td>
<td>0.12</td>
<td>0.57</td>
<td>0.53</td>
<td>0.42</td>
<td>0.67</td>
<td>10^{-12}</td>
<td>2.46</td>
<td>3.4</td>
<td>10^{-12}</td>
<td>0.965</td>
<td>0.965</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS 0.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IS 0.74</td>
<td>50</td>
<td>0.045</td>
<td>0.38</td>
<td>0.54</td>
<td>-0.058</td>
<td>0.69</td>
<td>0.54</td>
<td>-0.12</td>
<td>2.6</td>
<td>0.34</td>
<td>10^{-12}</td>
<td>1.86</td>
<td>1.86</td>
<td></td>
</tr>
<tr>
<td>H IS</td>
<td>25</td>
<td>0.045</td>
<td>0.38</td>
<td>0.54</td>
<td>-0.058</td>
<td>0.69</td>
<td>0.54</td>
<td>-0.12</td>
<td>2.6</td>
<td>0.34</td>
<td>10^{-12}</td>
<td>1.86</td>
<td>1.86</td>
<td></td>
</tr>
<tr>
<td>I IS</td>
<td>50</td>
<td>0.027</td>
<td>0.45</td>
<td>0.54</td>
<td>-0.058</td>
<td>0.69</td>
<td>0.54</td>
<td>-0.12</td>
<td>2.6</td>
<td>0.34</td>
<td>10^{-12}</td>
<td>1.86</td>
<td>1.86</td>
<td></td>
</tr>
</tbody>
</table>

(continued)
A.3. Detailed parameterization of the inference method

Analysis of the simulated data sets with Migraine is done in two or three automatic iterations depending on the demographic scenario. For the first iteration, \( n_p \) parameters points, with replicates for one every thirty of them, are sampled from an initial large range of parameter values set to \( \theta = [0.001 - 100] \), \( D = [0.001 - 20] \), and \( \theta_{\text{anc}} = [0.001 - 600] \), for all simulations. For analyses under the GSM, the parameter \( p \) is sampled from an initial range of \( [0 - 0.9] \). Such large exploration ranges is not very efficient in terms of parameter space exploration but allows us to automatically obtain good likelihood surface inferences for all simulated scenarios without the need of manual tuning for each scenario. Likelihoods are then estimated for these first \( n_p (1 + 1/30) \) points, and a likelihood surface is inferred by Kriging. From this likelihood surface, \( n_p \) additional points, with again replicates for one every thirty of them, are computed inside a convex envelope including the whole \( P=0.001 \) confidence region to limit or extend the parameter space previously explored. For the second iteration, likelihoods are estimated for these additional \( n_p (1 + 1/30) \) points and a second likelihood surface is inferred from all \( 2n_p (1 + 1/30) \) points. When a third step is considered, the final likelihood surface is then inferred from all \( 3n_p (1 + 1/30) \) points. This iterative procedure, described in details in Rousset and Leblois (2012) and in the Migraine user manual, appears extremely efficient as it ensures that many points are sampled near the top of the likelihood surface. For this work, we chose \( n_p = 600 \) for models with three parameters (i.e. under the SMM), and \( n_p = 2,400 \) for analyses under the GSM with four parameters.

As discussed in the main text (Section 3.2 in the main text), IS distributions are more or less efficient depending on the level of “disequilibrium” in the demographic model (e.g. the rate of instantaneous population size change). For this reason, we choose to simulate a high number ancestral histories \( (n_H = 2,000) \) to obtain good CI coverage properties in almost all scenarios explored. Doing so, we probably lost an important amount of computation time when analyzing scenarios for which much smaller \( n_H \) values are sufficient for good likelihood inference. However, it allows us to avoid choosing the minimal number of histories compatible with each scenario. The first exception was for a few situations with very recent and strong contractions \( (\theta = 0.4, \theta_{\text{anc}} = 400, D \leq 0.25) \), for which up to 200,000 ancestral histories were considered. The efficiency of the de Iorio and Griffiths (2004a)’s IS algorithms for time-inhomogeneous demographic models is further discussed in the main text.

Finally, analyses of 200 data sets and the consideration a large number \( n_p \) of parameter points and/or multiple iterative analyses may be time consuming by IS when a large number of histories \( n_H \) have to be considered ( see. For this reason, we used the fastest PACanc2ID-likelihood approximation (see A.1 for details on PACanc2ID-likelihood) when testing the effect of population structure with a GSM mutational model because inference under models with 4 parameters necessitate to explore a much larger number of points than with 3 parameters (i.e. \( n_p = 2,400 \) vs 600 for each iteration, see above).

A.4. \( \hat{\pi} \)'s computation under a Generalized Stepwise Mutation Model (GSM)

The main computation step using DeIorio and Griffiths’ approach is to solve a system of recursive equations (eq.10 in de Iorio and Griffiths, 2004b) which gives the \( \hat{\pi} \)s (see above, p.2, for a definition) for a given sample configuration \( n \) as a function of the demographic and mutational parameters of the model considered. For a single population with a scaled
mutation rate $\theta = 2N\mu$ and for any mutation model represented by the matrix of allelic state transition $P$ with $K$ possible allelic states, and $\sum_j p_{ij} = 1$, the system of recursive equations is the following:

$$(n + \theta)\hat{\pi}(j|n) = n_j + \theta \sum_{i=1}^{K} p_{ij} \hat{\pi}(i|n); \quad \text{for all } j \in [1; K]$$

(2)

with $n = \{n_j\}$ for $j \in [1; K]$ and $n = \sum_{j=1}^{K} n_j$ the total number of genes in the sample. In de Iorio et al. (2005), we considered an unbounded strict stepwise mutation model (SMM) on $i \in \{-\ldots,-2,-1,0,1,2,\ldots\}$ with $P_{ij} = 0.5$ if $|i - j| = 1$ and zero otherwise. We solved the above system of equations using Fourier transforms (eq. 3.5 to 3.9 in de Iorio et al., 2005), and obtained the following solution:

$$\hat{\pi}(j|n) = \sum_{k=-\infty}^{\infty} \frac{n_k}{n + \theta} c_{|k-j|} \left( \frac{\theta}{n + \theta} \right)$$

(3)

where

$$c_\ell(\rho) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\ell \xi) \frac{1}{1 - \rho \cos(\xi)} d\xi$$

$$= \frac{1}{\sqrt{1 - \rho^2}} \left[ \frac{\rho}{1 + \sqrt{1 - \rho^2}} \right]^\ell.$$

(4)

In a computer implementation, considering an infinite number of possible allelic states is not really tractable. However, using eq. 3 with a sufficiently large number of alleles (e.g. $K > 200$) leads to efficient computations as well as almost perfect coverage properties of confidence intervals as shown in this paper.

The aim here is to extend the $\hat{\pi}$ computations to consider a generalized stepwise mutation model (GSM), in which each mutation event equally leads to the gain or the loss of $X$ repeats, with $X$ being geometrically distributed with parameter $g$ (named $p$ in the main text). Under these conditions, $P_{ij} = g^{|i-j|}$ for $|i - j| \neq 0$ and zero for $i = j$ and Fourier transforms can also be used to solve the system for an infinite number of alleles (i.e., $i \in \{-\ldots,-2,-1,0,1,2,\ldots\}$) as in de Iorio et al. (2005). The characteristic function of the geometric distribution is

$$u^*(\xi) = \frac{(1 - g)(\cos(\xi) - g)}{1 - 2g \cos(\xi) + g^2}.$$

Then equation (3.17) of de Iorio et al. (2005) for computation of the $\hat{\pi}$ under a two population model becomes here

$$\hat{\pi}(j|\alpha, n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ij\xi} |A(\xi)|^{-1} \left[ (n_\beta q_{\beta}^{-1} + m_\beta + \theta(1 - u^*(\xi)))q_{\alpha}^{-1}n_{\alpha} + m_\alpha q_{\alpha}^{-1}n_{\alpha} \right] d\xi.$$

(5)

where $\alpha$ and $\beta$ are indices representing the two populations, and

$$|A(\xi)|^{-1} = \frac{1}{\lambda_1 - \lambda_2} \left[ \frac{1}{\theta - \lambda_1} \frac{1}{1 - \rho_1 u^*(\xi)} - \frac{1}{\theta - \lambda_2} \frac{1}{1 - \rho_2 u^*(\xi)} \right].$$
Following de Iorio et al. (2005), we get

\[ \hat{\pi}(j | \alpha, n) = \frac{1}{\lambda_1 - \lambda_2} \sum_{k=-\infty}^{+\infty} \left[ a_1(k, j) - a_2(k, j) \right] \]

where

\[ a_i(k, j) = \left\{ g^{-1}_\alpha \left[ n_\beta g^{-1}_\beta + m_\beta + \theta \right] n_{\alpha k} + m_\alpha g^{-1}_\beta n_{\beta k} \right\} I_{k-j}(\rho_i) - \left\{ \theta g^{-1}_\alpha n_{\alpha k} \right\} J_{k-j}(\rho_i) \]

and

\[ I_\ell(\rho) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\cos(\ell \xi)}{1 - \rho u^*(\xi)} d\xi \quad \text{and} \quad J_\ell(\rho) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{u^*(\xi) \cos(\ell \xi)}{1 - \rho u^*(\xi)} d\xi. \]

We thus need to compute \( I_\ell(\rho) \) and \( J_\ell(\rho) \) integrals. For \( I_\ell(\rho) \), we have

\[
\frac{1}{1 - \rho u^*(\xi)} = \frac{1 + g^2 - 2g \cos \xi}{1 + g^2 + \rho g(1 - g) - (2g + \rho(1 - g)) \cos \xi} = \frac{1}{1 + g^2 + \rho g(1 - g)} \left( \frac{1 + g^2}{1 - R \cos \xi} - \frac{2g \cos \xi}{1 - R \cos \xi} \right),
\]

with

\[ R := \frac{2g + \rho(1 - g)}{1 + g^2 + \rho g(1 - g)}. \]  

Since \((1 - g)^2(1 - \rho) > 0\), we have \(0 < (1 - g)^2 - \rho(1 - g)^2\), hence \(2g + \rho(1 - g) < 1 + g^2 + \rho g(1 - g)\), and thus \(R < 1\).

Using linearity of integration, and

\[ 2 \cos \xi \cos(\ell \xi) = \cos((\ell + 1) \xi) + \cos((\ell - 1) \xi) \]

in (6) yields

\[ I_\ell(\rho) = \frac{(1 + g^2)c_\ell(R) - g \left( c_{\ell+1}(R) + c_{\ell-1}(R) \right)}{1 + g^2 + \rho g(1 - g)} \]

with \(c_\ell\) as previously defined (eq. 4).

To compute \(J_\ell(\rho)\), we have

\[
\frac{u^*(\xi)}{1 - \rho u^*(\xi)} = \frac{(1 - g)(\cos \xi - g)}{1 + g^2 + \rho g(1 - g) - (2g + \rho(1 - g)) \cos \xi} = \frac{1}{1 + g^2 + \rho g(1 - g)} \left( \frac{g(1 - g)}{1 - R \cos \xi} + \frac{(1 - g) \cos \xi}{1 - R \cos \xi} \right).
\]

Again, using

\[ 2 \cos \xi \cos(\ell \xi) = \cos((\ell + 1) \xi) + \cos((\ell - 1) \xi) \]

yields

\[ J_\ell(\rho) = \frac{-g(1 - g)c_\ell(R) + \frac{1 - g}{2} \left( c_{\ell+1}(R) + c_{\ell-1}(R) \right)}{1 + g^2 + \rho g(1 - g)} \]

\[ \text{(10)} \]
Write $R_1$ and $R_2$ for the parameter defined in (7) when $\rho = \rho_1$ and $\rho = \rho_2$ respectively. Using (8) and (10), we obtain

$$a_i(k, j) = \frac{1}{D_i}\left\{q_{\alpha}^{-1} n_{\alpha k} \left[ (n_{\beta} q_{\beta}^{-1} + m_{\beta} + \theta)(1 + g^2) + \theta g(1 - g) \right] c_{k-j}(R_i) - q_{\alpha}^{-1} n_{\alpha k} \left[ g(n_{\beta} q_{\beta}^{-1} + m_{\beta} + \theta) + \theta(1 - g)/2 \right] \left( c_{k-j+1}(R_i) + c_{k-j-1}(R_i) \right) + q_{\beta}^{-1} n_{\beta k} m_{\alpha} \left[ (1 + g^2) c_{k-j}(R_i) - g(c_{k-j+1}(R_i) + c_{k-j-1}(R_i)) \right]\right\}$$

with $D_i = 1 + g^2 + \rho_i g(1 - g)$. Finally,

$$a_i(k, j) = \frac{1}{D_i}\left\{q_{\alpha}^{-1} n_{\alpha k} \left[ (n_{\beta} q_{\beta}^{-1} + m_{\beta} + \theta)(1 + g^2) + \theta(1 + g) \right] c_{k-j}(R_i) - q_{\alpha}^{-1} n_{\alpha k} \left[ g(n_{\beta} q_{\beta}^{-1} + m_{\beta} + \theta) + \theta(1 + g)/2 \right] \left( c_{k-j+1}(R_i) + c_{k-j-1}(R_i) \right) + q_{\beta}^{-1} n_{\beta k} m_{\alpha} \left[ (1 + g^2) c_{k-j}(R_i) - g(c_{k-j+1}(R_i) + c_{k-j-1}(R_i)) \right]\right\}. \quad (11)$$

When $g = 0$, mutation model is exactly SMM. Setting $g = 0$ in (11) leads to

$$a_i(k, j) = q_{\alpha}^{-1} n_{\alpha k} \left[ (n_{\beta} q_{\beta}^{-1} + m_{\beta} + \theta) \right] c_{k-j}(\rho_i) - q_{\alpha}^{-1} n_{\alpha k} \left[ \theta/2 \left( c_{k-j+1}(\rho_i) + c_{k-j-1}(\rho_i) \right) \right] + q_{\beta}^{-1} n_{\beta k} m_{\alpha} \left[ c_{k-j}(\rho_i) \right]$$

since $D_i = 1$ and $R_i = \rho_i$ in this case. This is exactly (3.19) of de Iorio et al. (2005).

For a single population model, i.e. setting all migration terms $\{m_i\}$ to 0 and only keeping terms with $\alpha$ indices, the expression for $\hat{\pi}$ reduces to eq. 3 except that the $c_{\ell}(\rho)$ terms are replaced by

$$I_{\ell}(\rho) = \frac{(1 + g^2) c_{\ell}(R) - g\left( c_{\ell+1}(R) + c_{\ell-1}(R) \right)}{1 + g^2 + \rho g(1 - g)}, \quad (12)$$

with

$$R = \frac{2g + \rho(1 - g)}{1 + g^2 + \rho g(1 - g)} \quad (13)$$

and $c_{\ell}(\rho)$ defined in eq. 4. As expected, the higher $g$ is, the more slowly $\hat{\pi}$ values are decreasing for distant mutation under the GSM than under the SMM. The above expression are increasingly poor approximations of bounded, and practically usable, GSMs as $g$ increases.

For a single population model, an alternative is to use the approach of Stephens and Donnelly (2000) to solve the system of recursive equations. Noting $\hat{\mathbf{n}}$ the vector of $\hat{\pi}(j|\mathbf{n})$,
their approach is based on the following representation of system of recursive equations:

\[
\hat{\Pi} = \hat{\Pi}^t \cdot I = \frac{n}{n + \theta} + \frac{\theta}{n + \theta} \cdot \hat{\Pi}^t \cdot P
\]

\[
\hat{\Pi}^t (I - \frac{\theta}{n + \theta} P) = \frac{n}{n + \theta}
\]

\[
\hat{\Pi}^t = (I - \frac{\theta}{n + \theta} P)^{-1} \frac{n}{n + \theta}
\]

where \( I \) is the identity matrix of dimension \( K \).

Using matrix inversion computations to solve the system is straightforward, but it is efficient only for mutation models with a relatively small number of alleles and for time-homogeneous models. It is computationally highly demanding for time-inhomogeneous models, as for the case of a population with variable size considered here, because it requires matrix inversions at almost every step of the backward history reconstruction, when \( \frac{n}{q} + \theta \) changes.

Instead of computing \( (I - \frac{\theta}{n + \theta} P)^{-1} \) each time an element of \( \hat{\Pi} \) is required and \( \theta, n \) or \( q \) values have changed, it is more efficient to compute the eigenvalues and eigenvectors of \( P \) once at the beginning of a run (actually, only once for each new value of \( g \)). Then, each time \( \theta, n \) or \( q \) changes, we compute \( R \cdot (I - \frac{\theta}{n + \theta} \Lambda)^{-1} \cdot L \), where \( \Lambda \) is the diagonal matrix of eigenvalues of \( P \), \( R \) is the matrix of its right eigenvectors and \( L \) that of its left eigenvectors. Moreover, the eigen decomposition allows one to evaluate the required \( \hat{\pi}(j|n) \) only, not the full \( \hat{\Pi} \) vector.

More details are as follows. Let \( \lambda_k, R^{(k)} \) and \( L^{(k)} \) be the \( k \)th eigenvalue and the \( k \)th right and left eigenvectors of \( P \), respectively. Let \( \mathbf{a} = (a_j) \) be the vector of coordinates of \( \hat{\Pi} \) in the basis of the left eigenvectors, such that \( \hat{\Pi} = \sum_{l=1}^{K} a_l L^{(l)} \). Since \( RL = I \) for any matrix (symmetric or not), each \( a_l \) is recovered as \( \hat{\Pi} \cdot R^{(l)} \), and

\[
\hat{\pi}(j|n) = \sum_{l=1}^{K} (\hat{\Pi} \cdot R^{(l)}) L^{(l)}.
\]

Reconsidering the initial system (eq. 2), we can write

\[
\sum_{j=1}^{K} \hat{\pi}(j|n) R^{(m)}_j = \frac{1}{\frac{n}{q} + \theta} \left( \sum_{j=1}^{K} \frac{n_j}{q} R^{(m)}_j + \theta \sum_{j=1}^{K} \sum_{i=1}^{K} p_{ij} \hat{\pi}(i|n) R^{(m)}_j \right)
\]

\[
= \frac{1}{\frac{n}{q} + \theta} \left( \sum_{j=1}^{K} \frac{n_j}{q} R^{(m)}_j + \theta \sum_{i=1}^{K} \hat{\pi}(i|n) \sum_{j=1}^{K} p_{ij} R^{(m)}_j \right).
\]

Using eigenvectors and eigenvalues definitions, we have \( \sum_{j=1}^{K} p_{ij} R^{(m)}_j = \lambda_i \cdot R^{(m)}_i \), and
\[
\sum_{j=1}^{K} \hat{\pi}(j|\mathbf{n}) R_j^{(m)} = \frac{1}{\frac{n}{q} + \theta} \left( \sum_{j=1}^{K} \frac{n_j}{q} R_j^{(m)} + \theta \sum_{i=1}^{K} \hat{\pi}(i|\mathbf{n}) \lambda_j R_i^{(m)} \right) \\
= \frac{1}{\frac{n}{q} + \theta} \left( \sum_{j=1}^{K} \frac{n_j}{q} R_j^{(m)} + \theta \lambda_m \sum_{j=1}^{K} \hat{\pi}(j|\mathbf{n}) R_j^{(m)} \right)
\]

(16)

then

\[
\sum_{j=1}^{K} \hat{\pi}(j|\mathbf{n}) R_j^{(m)} \left(1 - \frac{\theta \lambda_m}{\frac{n}{q} + \theta} \right) = \sum_{j=1}^{K} \frac{n_j}{q} R_j^{(m)} = \frac{n \cdot R^{(m)}}{\frac{n}{q} + \theta}
\]

(17)

and

\[
\sum_{j=1}^{K} \hat{\pi}(j|\mathbf{n}) R_j^{(m)} = \frac{n \cdot R^{(m)}}{\frac{n}{q} + \theta \lambda_m}
\]

(18)

To solve eq. 18, and get each \(\hat{\pi}(j|\mathbf{n})\), we use \(\hat{\Pi} = \sum_{l=1}^{K} a_l L^{(l)}\) and eq. 15, and get

\[
\hat{\pi}(j|\mathbf{n}) = \sum_{l=1}^{K} a_l L^{(l)} = \sum_{l=1}^{K} \hat{\Pi} \cdot R^{(l)} L^{(l)}_j
\]

(19)
B. Comparison with MsVar

Table S2: Simulated data sets from Girod et al. (2011)

<table>
<thead>
<tr>
<th>$D$ ($T$)</th>
<th>$\theta$ ($N$)</th>
<th>$\theta_{anc}$ ($N_{anc}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025 (10)</td>
<td>0.4 (200)</td>
<td>4.0 (2000)</td>
</tr>
<tr>
<td>0.025 (10)</td>
<td>0.4 (200)</td>
<td>40.0 (200,000)</td>
</tr>
<tr>
<td>0.025 (10)</td>
<td>0.4 (200)</td>
<td>400.0 (200,000)</td>
</tr>
<tr>
<td>0.125 (50)</td>
<td>0.4 (200)</td>
<td>4.0 (2000)</td>
</tr>
<tr>
<td>0.125 (50)</td>
<td>0.4 (200)</td>
<td>40.0 (200,000)</td>
</tr>
<tr>
<td>0.125 (50)</td>
<td>0.4 (200)</td>
<td>400.0 (200,000)</td>
</tr>
<tr>
<td>0.25 (100)</td>
<td>0.4 (200)</td>
<td>4.0 (2000)</td>
</tr>
<tr>
<td>0.25 (100)</td>
<td>0.4 (200)</td>
<td>40.0 (200,000)</td>
</tr>
<tr>
<td>0.25 (100)</td>
<td>0.4 (200)</td>
<td>400.0 (200,000)</td>
</tr>
<tr>
<td>1.25 (500)</td>
<td>0.4 (200)</td>
<td>4.0 (2000)</td>
</tr>
<tr>
<td>1.25 (500)</td>
<td>0.4 (200)</td>
<td>40.0 (200,000)</td>
</tr>
<tr>
<td>1.25 (500)</td>
<td>0.4 (200)</td>
<td>400.0 (200,000)</td>
</tr>
</tbody>
</table>

In Girod et al. (2011), five data sets were simulated for each of the scenario with the SimCoal2 software (Laval and Excoffier, 2004) and analyzed with the MsVAR software. For each scenario, 5 data sets of $n_g = 100$ genes, genotyped at $n_\ell = 10$ unlinked microsatellite loci were analyzed.

For comparison with MsVAR, we reanalyzed all 5 data sets (100 genes genotyped at 10 loci) considered in each of the 12 contraction scenarios of Girod et al. (2011) described in Table S2. All results are presented in Fig. S1, S2, and S3. Note that MsVAR is implemented under a Bayesian Framework. We thus compare Bayes factors vs. LRT for bottleneck detection and HPD intervals vs. confidence intervals for parameter inference.

First, Migraine and MsVAR globally give similar results, with a small but clear advantage of Migraine over MsVAR both in terms of higher BDRs and smaller FEDRs. For parameter inference, it is more difficult to draw clear conclusions from our comparison of the two methods because of the small number of data sets analyzed for each scenario. Analyses of such a limited number of simulated data sets in Girod et al. (2011) was due to the large computation times of MsVAR.

Over all demographic scenarios, the stronger differences in terms of parameter inference are due (1) to differences in the detection of past contraction for some data sets and (2) different behavior of the two methods for demographic situations with strong and recent changes in population size. First, when a contraction is detected with Migraine but not with MsVAR, parameter inference with Migraine is as expected more precise than with MsVAR. This phenomenon can be seen on Fig. S1 for some data sets for $D = 0.025$, 0.125 and 0.25. Second, for strong and recent contractions, MsVAR and Migraine have both difficulties to correctly infer likelihood surfaces and often give biased estimations for some parameters as well as too narrow CI which often do not contain the simulated value (Fig. S3). On one hand, MsVAR shows important convergence issues for those scenarios, as shown in Girod et al. (2011), often underestimates $\theta_{anc}$ and gives too narrow CIs. On the other hand, for the same scenarios, IS algorithms implemented in Migraine are much less efficient than for other less extreme scenarios (see Section 3.2). Migraine results reported on Fig. S3 were obtained by simulation of $n_H = 200,000$ ancestral histories for $D = 0.125$ and 0.25, and $n_H = 2,000$ for all other cases. However, even with such large exploration of ancestral histories, Migraine still give biased estimates but, contrarily to MsVAR, it gives CIs that almost always contain the expected value.

To further compare inferences without considering those two phenomena, we focussed on combinations of data sets and parameters for which there is enough information to obtain reasonable estimations with both methods and for which there is no MCMC convergence nor IS inefficiency issues. In such situations, we note that inference of $D$ and $\theta_{anc}$ using...
MIGRAINE and MsVAR give similar results in terms of point estimates and CIs, (e.g., Fig. S2 for $\theta_{anc} = 40.0$ and $D \geq 0.125$, Fig. S3 for $\theta_{anc} = 400.0$ and $D = 1.25$). However, compared to MIGRAINE’s estimations, MsVAR gives slightly lower $\theta_{anc}$ estimates with CIs that more often do not contain the simulated value. Both methods also give similar results for inference of $\theta$ in terms of point estimates and upper bounds of CIs (e.g. Fig. S2 with $D = 1.25$), but MsVAR sometimes infer lower CIs bounds that are well below those obtained with MIGRAINE (e.g. data sets #2 and #4).
Figure S1: Comparison of the results obtained with Migraine (black) and MsVar (gray) for the analyses of the 20 contraction data sets from Girod et al. (2011) with \( \theta = 0.4 \), \( \theta_{\text{anc}} = 4.0 \) and \( D = \{0.025; 0.125; 0.25; 1.25\} \) (see Table S2). Horizontal lines indicate the parameter value used for the simulation. For each data set numbered from 1 to 5, point estimates are represented by a square and Migraine confidence intervals and MsVar credibility intervals are represented by vertical lines. Dotted lines indicate an infinite bound for Migraine confidence intervals. BDR: bottleneck detection rate, FEDR: expansion detection rate. NC: proportion of data sets for which MsVar did not converge after 3 months. See Girod et al. (2011) for details about MsVar analyses.
Figure S2: Comparison of the results obtained with Migraine (black) and MsVar (gray) for the analyses of the 20 contraction data sets from Girod et al. (2011) with $\theta = 0.4$, $\theta_{anc} = 40.0$ and $D = \{0.025; 0.125; 0.25; 1.25\}$ (see Table S2).
Figure S3: Comparison of the results obtained with Migraine (black) and MsVar (gray) for the analyses of the 20 contraction data sets from Girod et al. (2011) with $\theta = 0.4$, $\theta_{\text{anc}} = 400.0$ and $D = \{0.025; 0.125; 0.25; 1.25\}$ (see Table S2).
C. LRT-P-value cumulative distributions for all scenarios considered in the manuscript

Figure S4: case A

Figure S5: case B
Figure S6: case C

Figure S7: case D
Figure S8: case E

Figure S9: case F
Figure S10: case G

Figure S11: case H

Figure S12: case I

Figure S13: case J
Figure S18: case O

Figure S19: case P
Figure S20: case Q

Figure S21: case R
Figure S22: case S
Figure S23: case T
Figure S24: case 1

Figure S25: case 2
Figure S26: case 4

Figure S27: case 5

Figure S28: case 6

Figure S29: case 7
Figure S30: case 8

Figure S31: case 9

Figure S32: case 11

Figure S33: case 12
Figure S34: case 13

Figure S35: case 14

Figure S36: case 15

Figure S37: case 16
Figure S38: case 17

Figure S39: case 18

Figure S40: case 19

Figure S41: case 20
Figure S42: case 31

Figure S43: case 31
Figure S44: case 32

Figure S45: case 33
Figure S46: case 34

Figure S47: case 35
Figure S48: case 36

Figure S49: case 37
Figure S50: case 45

Figure S51: case 46
Figure S52: case 47

Figure S53: case 48
Figure S54: case 49

Figure S55: case 50
Figure S56: Cumulative distributions of LRT-Pvalues for a recent and very strong contraction scenario, with $\theta = 0.4$, $D = 0.25$ and $\theta_{\text{anc}} = 400.0$ with (a) $n_H = 2,000$ and (b) $n_H = 20,000$ and (c) $n_H = 200,000$.

References


