



**HAL**  
open science

## Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data

Raphaël Leblois, Pierre Pudlo, Néron Joseph, François Bertaux, Champak Reddy Beeravolu, Renaud Vitalis, François Rousset

► **To cite this version:**

Raphaël Leblois, Pierre Pudlo, Néron Joseph, François Bertaux, Champak Reddy Beeravolu, et al.. Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data. *Molecular Biology and Evolution*, 2014, 31 (10), pp.2805-2823. 10.1093/molbev/msu212 . hal-02087423

**HAL Id: hal-02087423**

**<https://hal.science/hal-02087423>**

Submitted on 14 Jun 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Maximum-Likelihood Inference of Population Size Contractions from Microsatellite Data

Raphaël Leblois,<sup>\*,1,2,3</sup> Pierre Pudlo,<sup>1,3,4</sup> Joseph Neron,<sup>2</sup> François Bertaux,<sup>2,5</sup> Champak Reddy Beeravolu,<sup>1</sup> Renaud Vitalis,<sup>1,3</sup> and François Rousset<sup>3,6</sup>

<sup>1</sup>INRA, UMR 1062 CBGP (INRA-IRD-CIRAD-Montpellier Supagro), Montpellier, France

<sup>2</sup>Muséum National d'Histoire Naturelle, CNRS, UMR OSEB, Paris, France

<sup>3</sup>Institut de Biologie Computationnelle, Montpellier, France

<sup>4</sup>Université Montpellier 2, CNRS, UMR I3M, Montpellier, France

<sup>5</sup>INRIA Paris-Rocquencourt, BANG Team, Le Chesnay, France

<sup>6</sup>Université Montpellier 2, CNRS, UMR ISEM, Montpellier, France

\*Corresponding author: E-mail: raphael.leblois@supagro.inra.fr.

Associate editor: Asger Hobolth

## Abstract

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology. In this 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 (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 with alternative ones. Moreover, our implementation of a GSM allows more accurate analysis of microsatellite data, as we show that the violations of a single step mutation assumption induce very high bias toward false contraction detection rates. However, our simulation tests also showed some limits, which most importantly are large computation times 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.

**Key words:** demographic inference, maximum likelihood, coalescent, importance sampling, microsatellites, bottleneck, population structure, mutation processes, population contraction.

## Introduction

Understanding the demographic history of populations and species is a central issue in evolutionary biology and molecular ecology, for example, 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; Keller and Waller 2002; Frankham et al. 2006; 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 et al. 2002; Schwartz et al. 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 et al. 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 et al. 2000 and Peery et al. 2012). But they estimate neither the severity of the contraction 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 1994a; Emerson et al. 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, Ohta and Kimura 1973). In a recent study, Girod et al. (2011) evaluated the performance of *MSVAR* by simulation. They have shown that *MSVAR* clearly outperforms moment-based methods to detect past changes in population sizes, but appears only moderately robust to misspecification of the mutational model: Deviations from the SMM often induce “false” contraction detections on simulated samples from populations at equilibrium. Chikhi et al. (2010) also found a strong confounding effect of population structure on contraction detection using *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.

This work extends the importance sampling (IS) class of algorithms (Stephens and Donnelly 2000; de Iorio and Griffiths 2004a, 2004b) to coalescent-based models of a single isolated population with a unique past change in population size. Such a model is rather simple compared with complex demographic scenarios occurring in natural populations but inferences based on it can easily be tested by simulation and compared with existing methods. Furthermore, we provide explicit formula for a generalized stepwise mutation model (GSM; Pritchard et al. 1999), following de Iorio et al. (2005).

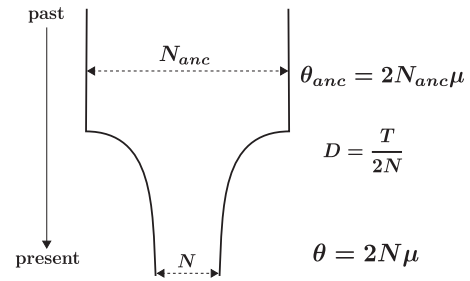
We have conducted three simulation studies to test the efficiency of our methodology on past contractions (i.e., bottlenecks) and its robustness against misspecifications of the model. The first study aims at showing the ability of the algorithm to detect contractions and to recover the parameters of the model (i.e., the severity of the population size change and its age) on a wide range of contraction scenarios. In the second study, we compared the accuracy of our IS implementation with the Monte Carlo Markov Chain (MCMC) approach implemented in *MSVAR*. The third study tests the robustness of our method against misspecification of the mutation model, and against the existence of a population structure not considered in the model. Finally, we have applied our methodology on the orangutan data set of Goossens et al. (2006) and compared our results with those obtained with *MSVAR*. All analyses in these studies were performed using the latest implementation of the *MIGRAINE* software package, available at <http://kimura.univ-montp2.fr/~rousset/Migraine.htm> (last accessed July 28, 2014).

## New Approaches

Our goal is to obtain maximum-likelihood (ML) estimates for single population models with a past variation in population size as described in the next section. To this end, we describe hereafter the successive steps of the inference algorithm.

### Demographic Model

We consider a single isolated population with a unique past size change (fig. 1). The method and our implementation in



**FIG. 1.** Representation of the demographic model used in the study.  $N$  is the current population size,  $N_{anc}$  is the ancestral population size (before the demographic change),  $T$  is the time measured in generation since present, and  $\mu$  is the mutation rate of the marker used. Those four parameters are the canonical parameters of the model.  $\theta$ ,  $D$ , and  $\theta_{anc}$  are the inferred scaled parameters.

*MIGRAINE* are quite general, in the sense that discrete (i.e., sudden), linear or exponential population size contractions or expansions can be considered. However, in agreement with Girod et al. (2011), we found in preliminary tests that parameter inference is less precise for expansions, especially for the time parameter. For this reason, we focused on contraction scenarios to test our method on smallish data sets with reasonable computation times (but see the Discussion section and supplementary fig. S5, Supplementary Material online, for the analysis of an expansion scenario). 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 function until reaching an ancestral population size  $N_{anc}$  at time  $t = T$ . Then,  $N(t) = N_{anc}$  for all  $t > T$ . More precisely,

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

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

### Computation of Coalescent-Based Likelihood with IS

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, that is, genealogies with mutations, leading to the current genetic data. Following Stephens and Donnelly (2000) and de Iorio and Griffiths (2004a), the Monte Carlo scheme computing this integral is based here on IS. The set of possible past

histories is explored through an importance distribution depending on the demographical scenario and the parameter values. The best proposal distribution to sample from is the importance distribution leading to a zero variance estimate of the likelihood. Here this distribution would be the distribution of gene history conditional on the current 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. However, it does not lead to a zero variance estimate of the likelihood. Better efficiency of the IS proposals allows to accurately estimate likelihoods by considering fewer histories for a given parameter value. Stephens and Donnelly (2000), de Iorio and Griffiths (2004a, 2004b), 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, 2004b) is the time inhomogeneity induced by the disequilibrium of our demographic model. Demographic models considered in de Iorio and Griffiths (2004a, 2004b) and in Rousset and Leblois (2007, 2012) suppose equilibrium and 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 all time-dependent quantities. To see how this is done, consider that the genealogy has been constructed until time  $T_k$ , the time of occurrence of the  $k$ th event, and that, at this date,  $n$  ancestral lineages remain. Under the coalescent with mutations, the expected rate of a mutation event is then  $n\theta/2$ , and  $n(n-1)\lambda(t)/2$  for a coalescence, where  $\lambda(t) = N/N(t)$  is the relative population size function describing demographic disequilibrium.  $\lambda(t)$  corresponds to parameter  $1/q$  in de Iorio and Griffiths (2004b). The total jump rate (i.e., occurrence rate of some event) at time  $t \geq T_k$  is then

$$\Gamma(t) = n((n-1)\lambda(t) + \theta)/2$$

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

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

Apart from these modifications that follow from the work of Griffiths and Tavaré (1994b), the outline of the IS scheme

from de Iorio and Griffiths (2004b) is preserved (see section A1 in the [supplementary material, Supplementary Material online](#), for more details).

We also develop specific algorithms to analyze data under the 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 an 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 if the range of allelic state is finite as soon as  $p$  is not very small (e.g.,  $<0.1$ ). To consider a more realistic GSM with allelic 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-dependent, this last decomposition is performed only once for a given matrix  $P$ . See section A4 in the [supplementary material, Supplementary Material online](#), for details about the GSM implementation.

Finally, several approximations of the likelihood, using products of approximate conditional likelihoods (Cornuet and Beaumont 2007) once the ancestral stable population is reached, and analytical computation of the probability of the last pair of genes, have been successfully tested to speed up computation times (see section A2 in the [supplementary material, Supplementary Material online](#)). In what follows, all analyses considering a GSM use these approximations, unless otherwise specified.

## Inference Method

Following Rousset and Leblois (2007, 2012), we first define a set of parameter points through 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 through 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 ML estimate. Finally, one- and two-dimensional profile likelihood ratios are computed, to obtain confidence intervals (CI) and graphical outputs (e.g., [fig. 2](#)). Section A3 in the [supplementary material, Supplementary Material online](#), explains how we tuned the parameters of the algorithm, namely the range

**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 an SMM, a GSM with  $p = 0.22$  and  $p = 0.74$ , a KAM and Two Situations with Variable Mutation Processes as Described in the Materials and Methods Section.

Case	$n_l$	$p$			$\theta$			$D$			$\theta_{anc}$			CDR (FEDR)
		Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
<b>SMM</b>														
[0]	10	NA	NA	NA	0.035	0.56	0.056	0.062	0.27	0.068	0.046	0.47	0.46	1 (0)
[A]	25	NA	NA	NA	0.0066	0.31	0.35	0.0079	0.16	0.73	0.0055	0.30	0.84	0.986 (0)
[B]	50	NA	NA	NA	0.015	0.23	0.62	0.0016	0.12	0.69	-0.0071	0.22	0.51	0.982 (0)
<b>GSM 0.22</b>														
[C]	10	0.26	0.91	0.16	0.033	0.51	0.12	0.16	0.66	0.14	0.22	1.33	0.657	0.990 (0)
[D]	50	0.17	0.47	0.12	0.059	0.25	0.44	0.012	0.14	0.75	-0.085	0.39	0.082	1.0 (0)
<b>GSM 0.74</b>														
[E]	10	0.016	0.14	0.0094	0.137	0.52	0.11	0.42	0.67	$< 10^{-12}$	2.46	3.4	$< 10^{-12}$	0.965 (0)
[F]	50	0.045	0.081	$3.8 \times 10^{-5}$	0.34	0.44	$< 10^{-12}$	0.40	0.49	$< 10^{-12}$	1.6	2.4	$< 10^{-12}$	1.0 (0)
<b>KAM</b>														
[G]	10	NA	NA	NA	-0.070	0.64	0.011	0.14	0.71	0.000034	2.11	4.8	0.012	0.84 (0)
[H]	25	NA	NA	NA	-0.027	0.49	0.54	-0.058	0.69	0.54	0.61	2.6	0.041	0.97 (0)
[I]	50	NA	NA	NA	-0.084	0.32	0.085	-0.22	0.51	0.19	0.402	2.74	0.0675	1.0 (0)
<b>var. mut. processes</b>														
[J]	10	0.18	0.91	0.00070	0.12	0.65	$9.8 \times 10^{-5}$	0.31	0.92	0.020	0.67	2.3	0.014	0.96 (0)
[K]	50	0.097	0.49	0.020	0.083	0.27	0.0055	0.040	0.18	0.99	-0.22	0.45	$4.3 \times 10^{-7}$	0.97 (0)

NOTE.— $n_l$ , number of loci; Rel. Bias, relative bias; KS,  $P$  value of the Kolmogorov–Smirnov test for departure of ECDF of LRT  $P$  values from uniformity; CDR, contraction detection rate; FEDR, false expansion detection rate; RRMSE, relative root mean square.

**Table 2.** Effects of Scaling the Time by the Mutation Rate Instead of Population Size for Different Timings,  $\theta = 0.4$  and  $\theta_{anc} = 40.0$ .

True $D$ or $D'$	Case	Scaling	$\theta$			$D$ or $D'$			$\theta_{anc}$		
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS
0.125 (0.05)	[3]	$D = T/2N$	2.6	5.7	$< 10^{-12}$	0.30	0.65	$2.3 \times 10^{-4}$	0.040	0.24	0.21
	[17]	$D' = T\mu$	2.3	4.5	$1.4 \times 10^{-9}$	2.6	5.61	$3.6 \times 10^{-10}$	0.0045	0.26	0.016
1.25 (0.5)	[0]	$D = T/2N$	0.035	0.56	0.056	0.062	0.27	0.068	0.046	0.47	0.46
	[18]	$D' = T\mu$	0.053	0.54	0.056	0.14	0.82	0.127	-0.0026	0.46	0.857
3.5 (1.4)	[7]	$D = T/2N$	-0.026	0.38	0.82	0.0038	0.50	0.51	0.32	1.7	0.098
	[19]	$D' = T\mu$	-0.013	0.37	0.91	0.020	0.71	0.12	0.389	2.11	0.40
5.0 (2.0)	[8]	$D = T/2N$	-0.107	0.36	0.33	-0.11	0.42	0.58	0.46	2.4	0.46
	[20]	$D' = T\mu$	-0.088	0.31	0.50	-0.16	0.52	0.68	0.49	2.5	0.60

NOTE.—Computations are done considering only data sets with a significant contraction detection. No effect of such scaling is detected on CDRs nor on FEDRs. Rel. Bias, relative bias; KS,  $P$  value of the Kolmogorov–Smirnov test for departure of ECDF of LRT  $P$  values from uniformity.

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.,  $N = N_{anc}$ ) and alternatives such as a population decline or expansion (i.e.,  $N \neq N_{anc}$ ). At level  $\alpha$ , our test rejects the null hypothesis if and only if 1 lies outside the  $1 - \alpha$  CI of the ratio  $N_{ratio} = N/N_{anc}$ .

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 is given in the Materials and Methods section.

## Results

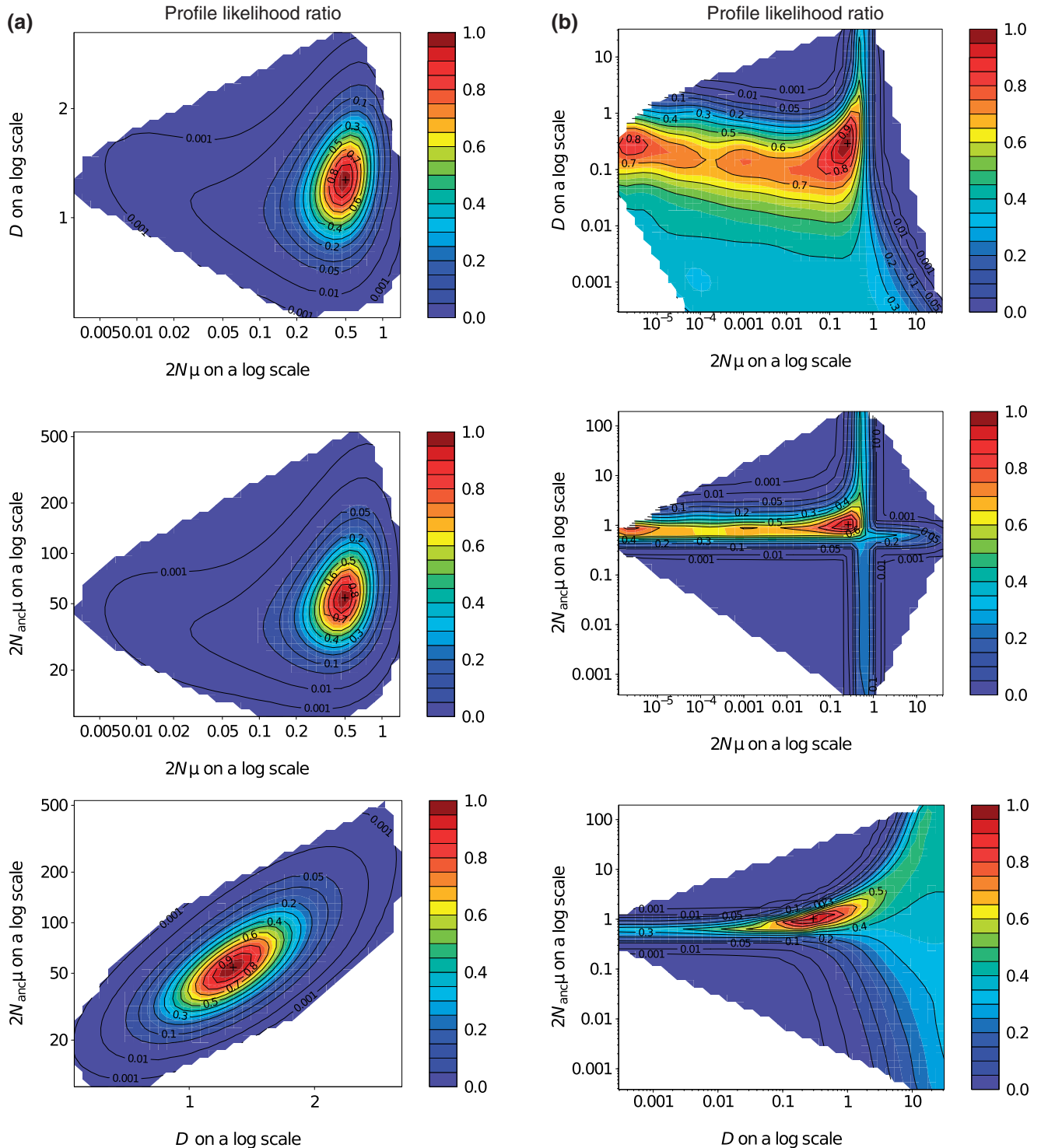
### Two Contrasting Examples

We first draw a contrast between two typical simulation outputs (figs. 2a and b), which must be kept in mind to understand further simulation results. The first one (case [0]), corresponding to our baseline simulation ( $\theta = 0.4$ ,  $D = 1.25$ , and  $\theta_{anc} = 40.0$ ), 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 ML values. The contraction signal is highly significant and is clearly seen in the  $(\theta, \theta_{anc})$  plot on figure 2a, as the ML peak is above the 1:1 diagonal. The second example (case [10]) is a more difficult situation, where the population has undergone a much weaker contraction ( $\theta = 0.4$ ,  $D = 1.25$ , and  $\theta_{anc} = 2.0$ ) 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 contraction signal is visible on the cross-shaped  $(\theta, \theta_{\text{anc}})$  plot on figure 2b, but is not significant.

### Implementation and Efficiency of IS on Time-Inhomogeneous Models

Simulation tests show that our implementation of de Iorio–Griffiths’ IS algorithm for a model of a single population with past changes in population size and stepwise mutations is



**FIG. 2.** Examples of typical two-dimensional profile likelihood ratios for two data sets generated with (a)  $\theta = 0.4$ ,  $D = 1.25$ ,  $\theta_{\text{anc}} = 40.0$  (case [0]) and (b)  $\theta = 0.4$ ,  $D = 1.25$ ,  $\theta_{\text{anc}} = 2.0$  (case [10]). The likelihood surface is inferred from 1,240 points in two iterative steps (a), and 3,720 points in three iterative steps (b) as described in section A3 in the [supplementary material, Supplementary Material](#) online. The likelihood surface is restricted to the region of the parameter space where the likelihood was actually estimated.

very efficient under most demographic situations tested here. Similar results are obtained for two different approximations of the likelihood (see section A2 in the [supplementary material, Supplementary Material](#) online). First, computation times are reasonably short: For a single data set with hundred gene copies and ten loci, analyses are carried out within few hours to 3 days on a single processor, even for the longest analyses with four parameters under the GSM. Second, likelihood ratio test (LRT)  $P$  value distributions generally indicate good CI coverage properties (see the Materials and Methods section). Empirical cumulative distribution functions (ECDF) of LRT  $P$  values for all scenarios, shown in section F in the [supplementary material, Supplementary Material](#) online, are most of the time close to the 1:1 diagonal as shown in [figure 3a](#) for our baseline scenario.

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  $P$  value properties were not always reached (e.g., [fig. 3b](#)) because of the small number of loci (i.e., 10) considered. Analyzing 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 the New Approaches section and section A1 in the [supplementary material, Supplementary Material](#) online. Contrarily to the first situation, likelihood surfaces are then too much peaked, and MLs 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 A1 in the [supplementary material, Supplementary Material](#) online) 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  $\theta_{\text{anc}} = 400.0$ ), increasing the number of ancestral histories sampled for each point up to 200,000 decreases relative bias and relative root mean square error (RRMSE) on  $\theta$  but does not provide satisfactory CI coverage properties ([supplementary fig. S64, Supplementary Material](#) online). Increasing the number of loci decreases the bias and RRMSE for  $D$  and  $\theta_{\text{anc}}$  but not for  $\theta$ . Such results have however only been observed in those few extreme situations with  $\theta/\theta_{\text{anc}} \leq 0.001$  and  $D \leq 0.25$ . [Figure 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  $\theta_{\text{anc}} = 40.0$ ).

ECDF of LRT  $P$  values for all parameters also more often depart 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  $p = 0.74$ , where  $p$  is the parameter of the geometric

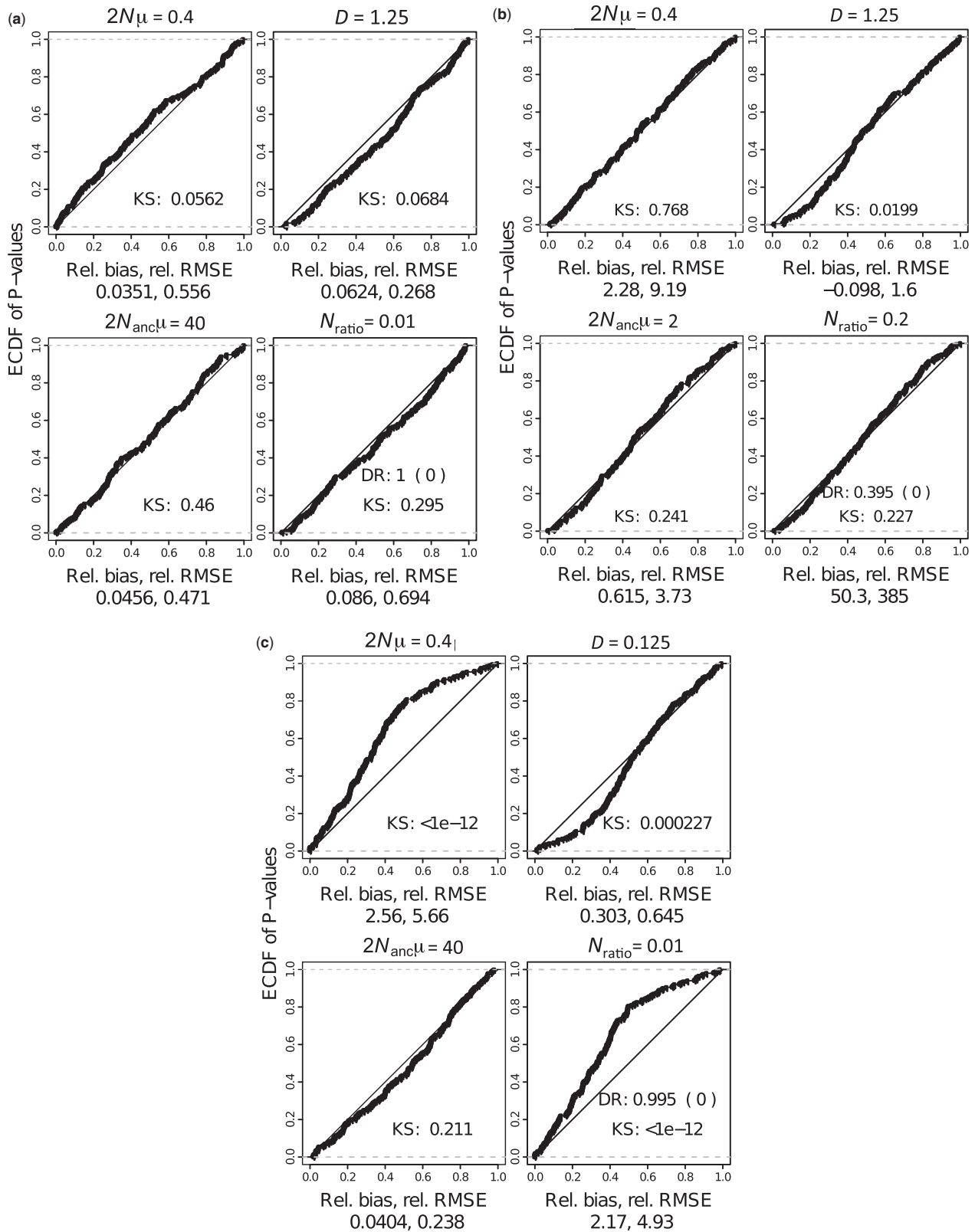
distribution of mutation step sizes, and for a K-allele model [KAM]: See [table 1](#), cases [E], [G], and [H]). In those situations, ECDF of LRT  $P$  values ([supplementary figs. S10, S12, and S13, Supplementary Material](#) online) indicate slightly too narrow CI, especially for parameters for which there is not much information (e.g.,  $\theta_{\text{anc}}$  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  $P$ -value distributions for case [I] but not for [F]: See [table 1](#) and [supplementary figs. S11 and S14, Supplementary Material](#) online). This suggests that the above incorrect ECDF of LRT  $P$  values are partly due to the small amount of information carried by a low number of loci but also due to slight misspecifications of the mutation model (i.e., the number of possible allelic states in the GSM, see section A2 in the [supplementary material, Supplementary Material](#) online).

### Power and Precision under Ideal Conditions

Results for the power of the contraction detection test and for the precision of the estimates under ideal conditions (i.e., same model used for simulations and analyses) with ten loci are presented in [figures 4 and 5](#).

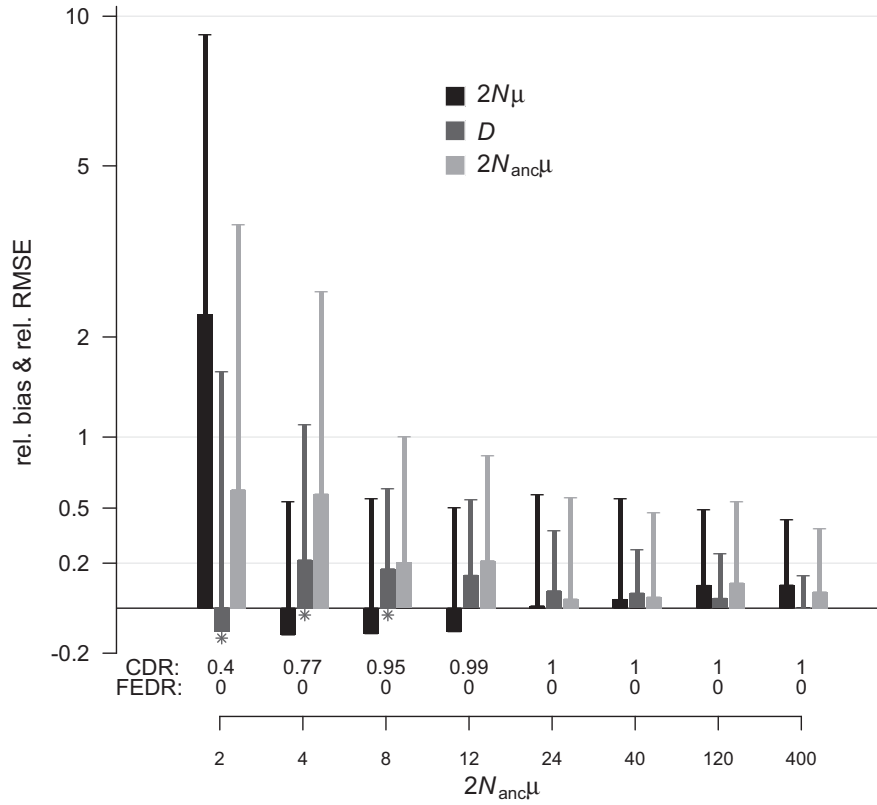
Contraction detection rates (CDRs) are highest when contractions are not too recent, nor too old or too weak: A contraction is detected at a 5% level in more than 95% of the data sets when the contraction occurred more than 25 generations ago but less than 1,400 generations ago ( $0.0625 < D < 3.5$ , [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 ( $\theta_{\text{anc}} = 5\theta$ , case [10], [fig. 4](#)) or extremely old ( $D = 7.5$ , case [9], [fig. 5](#)) contractions show CDRs 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 contraction. Reasonable precision, for example, a relative bias between  $-20\%$  and  $100\%$  and RRMSE below  $100\%$ , is only obtained when the ancestral population size is larger than 20 times the actual population size ([fig. 4](#)). However, for weaker contractions, estimates 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 [figure 5](#). Parameter  $\theta$  is inferred with good precision when the contraction is not too recent, for example, 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, for example, 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. Relative bias and RRMSE on  $D$  first decrease with time for contractions that occurred from 10 to 500 generations ago

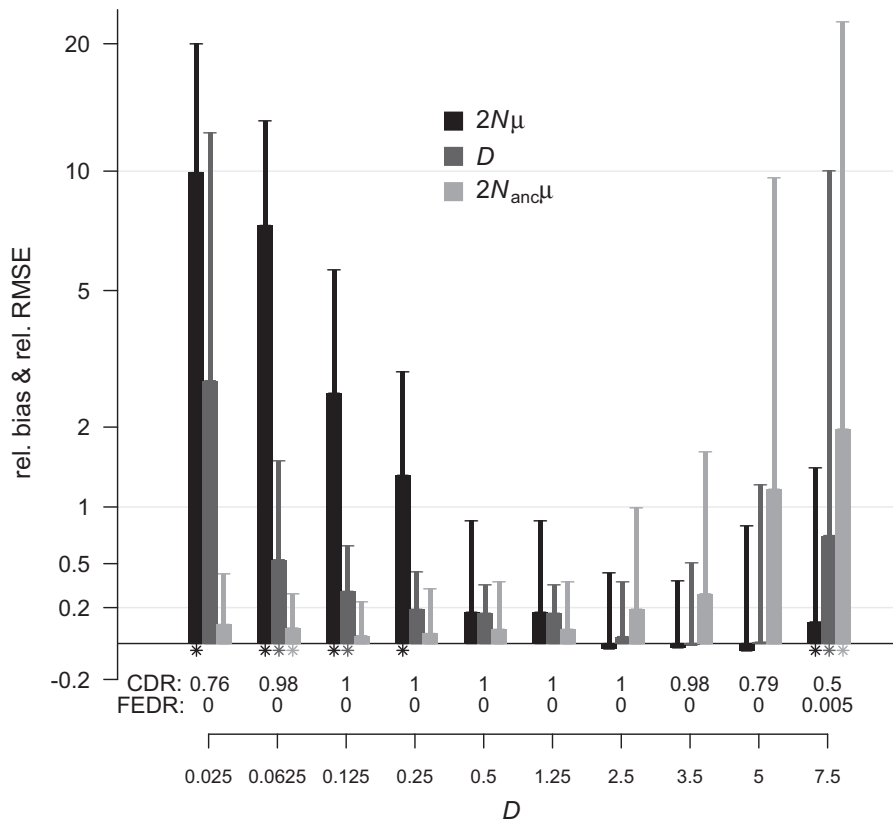


**FIG. 3.** ECDF of P values of LRTs for (a) the baseline scenario (case [0]), with  $\theta = 2N\mu = 0.4$ ,  $D = T/2N = 1.25$  and  $\theta_{\text{anc}} = 2N_{\text{anc}}\mu = 40.0$ ; (b) a very weak contraction scenario (case [10]), with  $\theta = 0.4$ ,  $D = 1.25$  and  $\theta_{\text{anc}} = 2.0$ ; and (c) a recent contraction scenario (case [3]), with  $\theta = 0.4$ ,  $D = 0.125$  and  $\theta_{\text{anc}} = 40.0$ . Mean relative bias (rel. bias, computed as  $\sum(\text{observed value} - \text{expected value})/\text{expected value}$ ) and relative root mean square error (rel. RMSE, computed as  $\sum[(\text{observed value} - \text{expected slope})/\text{expected value}]^2$ ) are reported as well as the contraction detection rate (DR) and FEDR in parentheses after DR. KS indicate the P value of the Kolmogorov-Smirnov test for departure of LRT P values distributions from uniformity.





**Fig. 4.** Effect of the strength of the population size contraction on the inference of each parameter of the model (cases [0] and [10]–[16]:  $2N\mu=0.4$ ). Relative bias is 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.,  $<0.05$ ). CDR, contraction detection rate; FEDR, false expansion detection rate.



**Fig. 5.** Effect of the timing of the population size contraction on the inference of each parameter of the model (cases [0]–[9]). See figure 4 for details.

( $0.025 < D < 1.25$ ), and then increase with time for older contractions.

Our baseline scenario (case [0]) with  $D = 1.25$ , thus seems to be the most favorable situation, for which inference of all parameters is relatively good given the small number of loci considered (ten loci; [figs. 3a, 4, and 5](#)). Relative biases are only about a few percent, but RRMSE values vary from 20% to 60% indicating different precision levels for the different parameters.  $D$  is the most precisely inferred parameter, followed by  $\theta_{\text{anc}}$  and then by  $\theta$ . The large RRMSE values are expectedly reduced when considering a larger number of loci, and reach 10–22% for all parameters when 50 loci are used ([table 1](#)).

A few simulations have been analyzed by inferring the parameter  $D' = T\mu$  instead of  $D = T/2N$ . For those simulations, we considered  $\theta = 0.4$  and  $\theta_{\text{anc}} = 40.0$  as in the baseline situation and four different timings ( $D = \{0.0125; 1.25; 3.5; 5.0\}$ , cases [17]–[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 CDRs nor on the false expansion detection rate (FEDR, results not shown).

### Effect of Mutational Processes

To test the robustness to mutational processes, we first analyzed under a strict SMM samples simulated under a stable population model with  $\theta = 2.0$  and a GSM with  $p = 0.22$  and  $0.74$  for the ten loci considered (cases [21] and [22]). For  $p = 0.22$ , 67% of the data sets show significant signals of false contraction. This false contraction detection rate (FCDR) increases up to 100% for  $p = 0.74$ . Among all simulations analyzed for these 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 contractions 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 (cases [23] and [24]).

Next, we simulated and analyzed data under a GSM or a KAM with a past contraction corresponding to our baseline scenario with  $\theta = 0.4$ ,  $D = 1.25$ , and  $\theta_{\text{anc}} = 40.0$ , and with  $p = 0.22$  and  $0.74$  for the GSM (with ten loci: Cases [C], [E] for the GSM and [G] for the KAM; with 50 loci: Cases [D], [F] for the GSM and [I] for the KAM, [table 1](#)). Compared with analyses under an SMM, CDRs slightly decrease when  $p$  increases but still remain very high (e.g.,  $\geq 95\%$  with ten 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 ECDF of LRT  $P$  values. However, increasing the number of loci from 10 to 50 restores good precision for the estimation of all

parameters, except for the KAM, as well as good LRT  $P$  value distributions. Finally, unaccounted variation in mutation processes and mutation rates across loci slightly increases biases and RRMSE, and induces poor CI coverage properties for the mutation parameters  $\theta$  and  $\theta_{\text{anc}}$  (cases [J] and [K] in [table 1](#), mutation processes detailed in the Materials and Methods section). The effect is similar but weaker for  $D$ , for which good precision and good CI coverage are observed with 50 loci.

### Effect of Population Structure

We first considered the presence of a local population structure by analyzing samples generated under stable continuous populations with various levels of dispersal and different spatial scale of sampling ([table 3](#)). All data sets were simulated under a GSM with  $p = 0.22$ . Our results show that isolation-by-distance (IBD) structure induces high FCDRs, strongly depending on the strength of IBD as well as the spatial scale of sampling (cases [25]–[29], [table 4](#)). The stronger the IBD structure is, the higher FCDR 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 ( $\sigma^2$  is the mean squared parent–offspring dispersal distance and is inversely related to the strength of IBD). Considering larger sampling scales by sampling on the whole population area not only strongly decreases FCDRs to values less than or equal to 16% for all levels of IBD but also induces false expansion detection in 4%, at most, of the data sets.

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 ([table 3](#)). We show in [table 5](#) that the presence of IBD slightly decreases CDRs, for example, 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  $\theta_{\text{anc}}$ , as well as bad CI coverage properties as shown by KS values ([table 5](#)) and ECDF of LRT  $P$  values ([supplementary figs. S46 and S48, Supplementary Material online](#)). Weaker IBD structure shows similar but weaker effects ([supplementary figs. S50 and S52, Supplementary Material online](#)). Increasing sample scale increases CDRs for situations under very strong IBD only, but strongly decreases relative biases and RRMSE on all parameters except  $\theta_{\text{anc}}$ . For all situations and all parameters, considering a large sample scale allows better CI coverage properties ([table 5](#) and [supplementary figs. S47, S49, S51, and S53, Supplementary Material online](#)).

We finally tested the influence of an island population structure with varying levels of migration and population sizes ([tables 6 and 7](#) and [supplementary tables S3 and S4, Supplementary Material online](#)). Our results first show that sampling a single island from a stable-structured population induces high FCDRs, from 11% to 52% depending on the level of population structure. With such local sampling scheme, the relationship between FCDRs 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 FCDRs up to 100%

**Table 3.** 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 through truncated Pareto distributions with mean squared parent-offspring dispersal distance, say  $\sigma^2$ , varying in {1; 4; 10; 20; 100}.

**Parameters of the IBD Populations:**

- At equilibrium:  $\theta = 4.0$  with a  $32 \times 31$  lattice (hence  $N = 1,984$  genes)
- Including an habitat contraction:  $(D, \theta, \theta_{\text{anc}}) = (1.25, 0.4, 40.0)$  with lattices of sizes from  $10 \times 10$  ( $N = 200$ ) to  $100 \times 100$  ( $N_{\text{anc}} = 20,000$ ) 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 four nodes) [large sample scale].

**Island Population Structure**

We considered models with  $d = 10$  demes of equal size  $N_d$  genes, varying in {20; 200; 1,000; 2,000}, and exchanging migrants at rate  $m$  between pairs of demes, varying in {0.00025; 0.00025; 0.0025; 0.005; 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  $D = 1.25$ ,  $\theta_{\text{anc}} = 40.0$

**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].

**Table 4.** Effects of IBD on the Detection of False Contraction and Expansion Signals in Constant-Size Populations with  $\theta = 4.0$ .

IBD Strength ( $\sigma^2$ )	Case	Sampling Scale	
		Small ( $10 \times 5$ ) FCDR/FEDR	Large ( $28 \times 28$ ) FCDR/FEDR
1	[25]	0.67/0.0	0.16/0.005
4	[26]	0.54/0.0	0.095/0.010
10	[27]	0.49/0.0	0.080/0.010
20	[28]	0.41/0.005	0.090/0.020
100	[29]	0.145/0.005	0.11/0.040

NOTE.—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 is inversely related to the strength of IBD. See the Materials and Methods section for details.

for highly structured situations (i.e.,  $M \leq 1.0$ , with  $M \equiv 2N_d m$ ); but it also reduces FCDRs for less structured populations, down to values around 10% for the larger sampling scale. Interestingly, sampling a single gene per deme strongly reduces the effect of population structure and decreases FCDR values to 2% in a situation with intermediate migration rates (i.e.,  $M = 1.0$ , [supplementary table S3, Supplementary Material online](#)). Note that a few false expansions are also detected among all those simulations but always in less than 10% of the data sets. Finally, we can also note that, as shown in [table 6](#) and [supplementary table S3, Supplementary Material online](#), increasing the total diversity  $\theta$  or considering an SMM also modifies the effect of population structure but more simulations are needed to find global trends for effect of genetic diversity and mutational processes.

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 7](#) and [supplementary table S4, Supplementary Material online](#)). Compared with unstructured situations with  $\text{CDR} = 99\%$  (case [G]), sampling a single deme strongly reduces CDRs 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 leads to small CDR 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 CI ([table 7](#) and [supplementary figs. S54–S57, Supplementary Material online](#)). This is observed whether the parameters considered are local values for one deme or global values for the whole population (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 ([supplementary figs. S58 and S59, Supplementary Material online](#)). With a single exception concerning samples with a single gene per deme ([supplementary table S4, Supplementary Material online](#)), increasing sampling scale generally increases CDRs. However, contrarily to the results obtained for IBD, sampling at a large scale, even with a single gene per deme, does not improve all parameter inferences nor all CI coverage properties. Sampling at a larger scale seems to often allow better estimation of  $\theta_{\text{anc}}$ , but the effect on all other parameters is highly dependent on the demographic scenario considered, and no clear conclusion can thus be drawn from our simulations.

**Table 5.** Effects of IBD Structure on the Detection and Characterization of a Past Contraction.

IBD Level ( $\sigma^2$ )	Case	Sample Scale	$p$			$\theta$			$D$			$\theta_{anc}$			CDR(FEDR)
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
1	[30]	Small	0.71	1.2	$6.6 \times 10^{-9}$	-0.30	0.43	$1.4 \times 10^{-10}$	0.90	1.2	$<10^{-12}$	0.25	1.2	0.061	0.9 (0)
		Large	0.20	0.88	0.14	-0.0577	0.46	0.057	0.46	0.79	$<10^{-12}$	0.51	1.4	0.22	0.99 (0)
4	[32]	Small	0.50	1.1	$5.1 \times 10^{-9}$	-0.29	0.45	$2.7 \times 10^{-7}$	0.43	0.78	$6.3 \times 10^{-9}$	0.25	1.4	0.46	0.96 (0)
		Large	0.22	0.89	0.74	-0.12	0.49	0.020	0.27	0.54	$4.4 \times 10^{-4}$	0.37	1.3	0.93	0.96 (0)
10	[34]	Small	0.41	1.0	$1.4 \times 10^{-5}$	-0.19	0.42	$3.4 \times 10^{-6}$	0.39	0.72	$1.9 \times 10^{-6}$	0.40	2.15	0.0031	0.98 (0)
		Large	0.23	0.89	0.12	-0.11	0.44	0.15	0.22	0.52	$1.4 \times 10^{-4}$	0.31	1.2	0.33	0.97 (0)
100	[36]	Small	0.35	0.96	$1.6 \times 10^{-4}$	-0.094	0.41	0.11	0.26	0.55	$6.0 \times 10^{-4}$	0.22	1.2	0.71	0.97 (0)
		Large	0.19	0.86	0.40	-0.017	0.48	0.67	0.13	0.46	0.14	0.26	1.3	0.52	0.96 (0)

NOTE.—Samples are simulated from a single continuous population under IBD that has undergone a past contraction with  $\theta = 0.4$ ,  $D = 1.25$ , and  $\theta_{anc} = 40.0$ . The small sampling scale corresponds to 100 genes sampled on a  $5 \times 10$  area, expressed in lattice nodes, in the center of the population; the large sampling scale corresponds to a spatially homogeneous sample of 100 genes taken on the whole population area (i.e., one individual every four nodes). See the Materials and Methods section for details.

**Table 6.** Effects of an Island Population Structure on the Detection of False Contraction or Expansion Signals in Constant-Size Populations.

$\theta$	Island Model Settings		Case	Sampling Scale		
	$M$			Small One Island FCDR/FEDR	Large Three Islands FCDR/FEDR	Very Large All Ten Islands FCDR/FEDR
4	0.01		[38]	0.11/0.025	1.0/0.0	1.0/0.0
	0.1		[39]	0.32/0.02	1.0/0.0	1.0/0.0
	1.0		[40]	0.21/0.0	0.84/0.0	0.76/0.0
	10.0		[41]	0.52/0.0	0.32/0.0	0.10/0.010
	30.0		[42]	0.38/0.0	0.18/0.01	0.10/0.015
	100.0		[43]	0.19/0	0.085/0.015	0.11/0.026
20	1.0		[44]	0.78/0.0	0.91/0.0	0.64/0.0

NOTE.—Samples are simulated from a stable island model with  $n_d$  demes,  $\theta = 2n_d N_d \mu$ , and scaled migration rate  $M \equiv 2N_d m$ . Sampling scale corresponds to the number of sampled demes. See the Materials and Methods section for details. FCDR, false contraction detection rate; FEDR, False expansion detection rate.

**Table 7.** Effects of an Island Population Structure on the Detection and Characterization of a Past Contraction.

Gene Flow Level ( $M$ )	Case	Sampling Scale	$p$			$\theta$			$D$			$\theta_{anc}$			CDR (FEDR)
			Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	Rel. Bias	RRMSE	KS	
0.01	[45]	Small	-0.081	1.0	0.026	-0.61	0.80	$7.0 \times 10^{-10}$	-0.24	1.2	0.060	-0.99	1.0	$<10^{-12}$	0.005 (0.040)
		Very large	2.2	2.3	$<10^{-12}$	-0.77	0.81	$<10^{-12}$	-0.66	0.67	$<10^{-12}$	-0.17	0.62	$8.5 \times 10^{-4}$	1.0 (0)
1.0	[47]	Small	1.6	1.9	$<10^{-12}$	-0.32	0.83	$1.7 \times 10^{-5}$	0.41	1.8	$9.9 \times 10^{-8}$	-0.94	0.96	$<10^{-12}$	0.070 (0.070)
		Very large	-0.0027	0.61	$4.2 \times 10^{-8}$	-0.72	0.80	$<10^{-12}$	-0.60	0.60	$<10^{-12}$	-0.020	0.50	$5.7 \times 10^{-5}$	1.0 (0)
100	[49]	Small	0.69	1.2	$4.2 \times 10^{-6}$	-0.070	0.43	0.64	0.31	0.80	0.0085	0.043	1.1	0.25	0.88 (0)
		Very large	0.014	0.83	0.0085	0.13	0.55	0.013	0.041	0.39	0.033	0.24	1.1	0.78	0.99 (0)

NOTE.—Samples are simulated from a 10-island model in which each subpopulation has undergone a past contraction, with  $\theta \equiv 2dN_d \mu = 0.4$ ,  $D \equiv T/(2dN_d) = 1.25$ , and  $\theta_{anc} \equiv 2dN_{d,anc} \mu = 40.0$ , and varying scaled migration rate  $M \equiv 2N_d m$ . See the Materials and Methods section for details. Mean relative bias and RRMSE are reported as well as the CDR and the FEDR. Sampling scale corresponds to the number of sampled demes: 1) Small for one sampled deme and 2) very large for ten sampled demes. KS indicate the  $P$  value of the Kolmogorov–Smirnov test for departure of ECDF of LRT  $P$  values from uniformity.

### Inferences on the Orangutan Data Set

The orangutan analyses with MIGRAINE show consistent results for the three pooled samples RS1, RS2 and RS1+RS2 (table 8 and supplementary fig. S4, Supplementary Material online), and for each subsample separately (supplementary table S5, Supplementary Material online). With one exception for subsample S9, all analyses detect a strong and recent past contraction of population size and allow concordant estimation of the model parameters. However, as expected due to lower sample sizes, analyses of each subsample give less

precise inferences than for the pooled sampled, and we will thus focus on the pooled sample results. First,  $\theta_{anc}$ 's inference is extremely consistent across analyses, and shows a high precision level with point estimates around 7.5 and narrow CI (i.e., around [5 – 12]). Second, the time when the contraction started in the past,  $D$ , is inferred with slightly less precision but all analyses support a relatively recent contraction with upper bounds of CI below 1.0. Third, in agreement with our simulation results, there is much less information about  $\theta$  because the inferred contraction is recent. High  $\theta$  values are

**Table 8.** Point estimates and 95% CI for All Model Parameters Obtained from the Analyses of the Orangutan Data Set.

Sample (size)	$p$	$\theta$	$D$	$\theta_{\text{anc}}$	$N_{\text{ratio}}$
RS1 (106)	0.40 [0.15 – 0.61]	0.0048 [ $10^{-5}$ – 0.36]	0.30 [0.17 – 0.80]	7.7 [5.2 – 11.5]	0.00063 [ $10^{-6}$ – 0.049]
RS2 (89)	0.42 [0.20 – 0.65]	0.00035 [ $10^{-5}$ – 0.41]	0.31 [0.19 – 0.48]	7.4 [5.4 – 9.9]	$4.8 \times 10^{-5}$ [ $1.2 \times 10^{-6}$ – 0.058]
RS1+RS2 (195)	0.37 [0.15 – 0.59]	0.013 [ $8 \times 10^{-5}$ – 0.67]	0.14 [0.045 – 0.52]	7.8 [5.3 – 11.8]	0.0016 [ $10^{-8}$ – 0.090]
		$N/2$	$T_{\text{years}}$	$N_{\text{anc}}/2$	
RS1		3 [1 – 180]	90 [17 – 14,400]	3,850 [2,600 – 5,750]	
RS2		1 [1 – 205]	31 [19 – 9,840]	3,700 [2,700 – 4,950]	
RS1+RS2		7 [1 – 335]	98 [5 – 17,420]	3,900 [2,650 – 5,900]	

NOTE.—More detailed results and a figure showing profile likelihood ratios are available in section D in the [supplementary material, Supplementary Material](#) online. The lower part of the table presents the estimates of population sizes expressed as numbers of diploid individuals ( $N/2$  and  $N_{\text{anc}}/2$ ) and times in years ( $T_{\text{years}}$ ) obtained after a conversion of MIGRAINE results using a fixed mutation rate of  $5 \times 10^{-4}$  mutation per locus per generation and a generation time of 25 years. The “confidence” intervals reported for  $T_{\text{years}}$  is likely to be much larger than the true 95% CI, because we used the 95% CI bounds of  $D$  and  $\theta$  successively to compute this interval. Sample sizes are given in number of diploid individuals. See the Materials and Methods section for details.

rejected, with upper bounds of CI between 0.4 and 0.7, but the likelihood profile is always very flat for low  $\theta$  values (e.g., [supplementary fig. S4, Supplementary Material](#) online). As a result, many ML estimates were inferred at the lower bound of the explored parameter space for  $\theta$  in preliminary analyses. Moreover, [supplementary figure S4, Supplementary Material](#) online, shows a clear trade-off between  $\theta$  and  $D$  for high likelihood with low  $\theta$  values associated with high values for  $D$ . Unrealistically, low  $\theta$  values (e.g.,  $10^{-9}$ ) were thus associated with high  $D$  values (e.g., about 1.0) in the first preliminary analyses. For the final analyses, we thus restricted the parameter space to biologically realistic  $\theta$  values (i.e.,  $\theta \geq 10^{-5}$ , as the latter value corresponds to a population size of a single individual when considering a very small mutation rate for microsatellite markers of  $2.5 \times 10^{-6}$ ). Because of this uncertainty in the estimation of  $\theta$ , inference of  $N_{\text{ratio}}$  clearly shows a population size ratio below 1, but the CI are very large, even when  $\theta$  values are constrained. Finally, the GSM parameter  $p$  is also inferred with intermediate precision but, in all analyses, its point estimates are relatively high, for example, around 0.4, compared with values generally found in the literature.

## Discussion

In this study, we adapted de Iorio–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 did not explore expansion scenarios because preliminary simulations showed that parameter inference is less precise for expansions than for contractions (as shown in section E in the [supplementary material and fig. S5, Supplementary Material](#) online, and in [Girod et al. 2011](#)). For a majority of expansion scenarios, good precision is only obtained when considering large sample sizes, for example, 500 haploid individuals genotyped at 50 loci, which implies large computation times (results not shown). Likewise, preliminary tests under a model with a

founder event followed by a demographic expansion showed that correct parameter inference could not be obtained with the current version of our IS algorithms within reasonable computation times. This is so because the strong disequilibrium of this model induces high variance of the likelihood estimation, which therefore requires the computation of a very large number of ancestral histories. We thus focused on a model with a single past contraction event to test the effect of the timing and amplitude of the past demographic change, and study the robustness of inferences to misspecifications of the mutational and population structure models. Our results allow us to illustrate both the strengths and the imperfections of the method.

## 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 contractions, e.g.,  $N_{\text{anc}}/N \geq 1,000$ ). For the later situations, much longer runs are needed to obtain good CI coverage. This shows that the efficiency of de Iorio–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. CDRs 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 ten 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 ability of the method to infer past demography depends on the genetic information available in the data and this information 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 occurred neither too recently nor too far in the past because in such scenarios coalescent and mutation 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 figures 4 and 5. One important result of our study is that high CDRs as well as good inference precision for both the time and the actual population size parameters are still expected for relatively ancient contraction (e.g.,  $1.25 \leq D \leq 5.0$ ).

Finally, many recent software packages 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), 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 simulations show that there is not much interest to scale time by mutation rate for inferences of past contractions. For the demographic scenarios considered here, such scaling always reduces inference precision for the time parameter. Beside those scaling issues, independent information about mutation rates of the markers 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.

### 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 contractions 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 CDRs, 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 contraction detection, parameter estimation, and computation times. For example, both methods perform well for intermediate contraction 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 shows strong convergence issues for very recent and strong contractions (see fig. 1 in Girod et al. 2011) and gives biased point estimates as well as bad CI for  $\theta_{\text{anc}}$  (supplementary fig. S3, Supplementary Material online). For the same demographic scenarios, IS algorithms implemented in MIGRAINE are not efficient and, even with large computation times, MIGRAINE shows high relative biases and RRMSE, as well as bad coverage properties of CI for  $\theta$  as discussed above. For both methods, computation times thus greatly increase with the strength of the contraction, and accurate parameter inference considering very recent and strong contractions may be difficult to achieve. However, MIGRAINE appears 1) more adapted to the analyses of microsatellite markers because of the implementation of the GSM, as detailed in the next section; 2) slightly more powerful than MSVAR as our simulations show higher CDRs (e.g., often more than 20% higher CDRs depending on the demographic situation considered) and fewer false expansion detections (i.e., 0 vs. 1 false expansion detected among all simulations with MIGRAINE and MSVAR, respectively); 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 number of available cores.

### 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, LAMARC, 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 an SMM (Garza and Williamson 2001). Peery et al. (2012)

recently showed that identification of past contractions using the summary statistic-based *BOTTLENECK* (Cornuet and Luikart 1996) and *M-RATIO* softwares (Garza and Williamson 2001) is highly biased by deviations from the mutation models implemented in those softwares, often leading to significant contraction detection in samples simulated under a stable population model. Faurby and Pertoldi (2012) also showed that estimation of current and past population sizes with *MSVAR* is unreliable when realistic deviations from the SMM occur. In the previous study of Girod et al. (2011), it was also shown that *MSVAR* was moderately robust to deviations from the SMM, which leads to false contraction detections in samples simulated from stable populations. However, this conclusion was presumably overoptimistic due to the small number of data sets analyzed. In this study, we clearly show a strong impact of violations of the SMM assumptions: Even small deviations from the SMM induce large FCDRs 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 first show that, in samples from stable populations, using a GSM successfully decreases the rate of false contraction detections due to mutations of more than one step. Second, for samples that effectively experienced a past contraction, our simulations show that using a GSM leads to CDRs similar to the one observed under the SMM, and also show that parameter inference precision is only slightly affected by the additional parameter  $p$ . 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 four 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. Finally, *MIGRAINE* considers that mutation rates are constant across loci and we showed that unaccounted variation in mutation processes across markers 1) increases estimation biases essentially when a low number of loci are considered (i.e., 10 vs. 50 loci); but 2) also slightly deteriorates CI coverage properties, principally for  $\theta$  and  $\theta_{anc}$  and regardless of the number of loci.

### 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, IBD 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 induces large FCDRs when individuals are sampled at a small spatial scale from stable populations, even for relatively weak IBD. However, sampling individuals at a larger scale, that is, over the whole population area, efficiently reduces FCDRs. Parameter estimation, and to a lesser extent CDRs, obtained from samples coming from a population that effectively went through a contraction is also affected by IBD, and again sampling individuals at a large geographical scale efficiently reduces the impact of IBD. Parameter inference thus appears robust unless IBD is very strong and sampling scale is 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 FCDRs unless gene flow is extremely limited. Considering larger sampling scales by sampling individuals from all the demes of the total structured population reduces FCDRs 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, that is,  $M \leq 1.0$ , often increases FCDRs. Our simulations finally show that, when a contraction did occur in the past, ignoring island population structure also often strongly decreases CDRs and greatly biases parameter estimation. Moreover, both contraction 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 CDRs below 10% and accurate CDRs are only obtained using large sampling scales. Parameter inference appears highly biased for all levels of gene flow considered in this study. As for CDRs, 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 current and past population sizes, and of the timing of the demographic change. Moreover, our results show complex interactions between levels of population structure, total genetic diversity, mutation processes, and sampling scales that strongly limit practical recommendations for the detection and characterization of past changes in population size in the presence of unaccounted population structure.

Such confounding effects of population structure and past changes in population sizes have already been observed. First, the effect of small-scale IBD population structure on CDRs obtained with the *BOTTLENECK* and *M-RATIO* softwares has been tested by simulations in Leblois et al. (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 nonnull 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 et al. (2010), Chikhi et al. (2010), and Heller et al. (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 contraction. 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 population 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 corresponds 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 (WF) 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 populations 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 oversimplified past genealogical processes may not always give the right prediction. Nevertheless, two 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 of past demographic changes only when migration rates are relatively high, that is,  $M \geq 10.0$ . Sampling a single gene per deme efficiently prevents false contraction detections in stable-structured populations but does not allow precise characterization of past contractions in the presence of intermediate to strong population structure; 2) for all other demographic situations, detection of past population size changes and parameter inferences based on panmictic models may often be misleading.

Such results finally imply that models themselves should be improved. First, model choice procedures 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 et al. (2010) used an Approximate Bayesian Computation 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., 60–85.5% with 10–200 loci, respectively). An alternative is to develop models accounting for both population structure and population size changes that 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.

### Analysis of the Orangutan Data Set

Our analyses of the orangutan data set show that 1) all sampled sites, except S9, exhibit a clear signal of a strong and recent population size contraction; and 2) parameter inferences are extremely consistent among sites, and among the different pooled samples. Those results are in good agreement with equivalent analyses using *MSVAR* published in Goossens et al. (2006) and Sharma et al. (2012): All analyses indicate 1) ancestral population sizes of about a few thousand individuals (i.e., [2,600–5,900] for *MIGRAINE* and [3,100–13,400] for *MSVAR*), 2) much smaller current population sizes (i.e., [1–335] individuals for *MIGRAINE* and [22–1,400] with median values between 60 and 200 individuals for *MSVAR*), and 3) a relatively recent timing of the contraction (i.e., less than about 15,000 years ago for both *MIGRAINE* and *MSVAR*). However, considering the same generation time of 25 years, *MSVAR* results show less support for a very recent event (i.e., <200 years) than *MIGRAINE* ones. For this reason, Sharma et al. (2012) conclude that *MSVAR* analyses suggest that the main historical factor explaining the inferred contraction is habitat destruction due to the arrival of farmers 4–5 ka, whereas we cannot exclude from our *MIGRAINE* analyses that habitat loss through more recent deforestation for the development of massive agriculture and logging in the last 150–200 years is also a likely cause of the decline of orangutan populations in Borneo.

As discussed in the above sections, the two major problems with such inference of past changes in population size are misspecification of the mutation processes and unaccounted population structure. First, *MIGRAINE* analyses used a combination of *GSM* and *SMM* models for di- and tetranucleotide microsatellite loci, respectively, because mutations at tetranucleotide loci have been shown to be principally single steps whereas dinucleotide markers show more multistep mutations (Sun et al. 2012). Note that this adequation between marker types (di- vs. tetranucleotides) and mutation models (*GSM* vs. *SMM*) was empirically validated in



preliminary simulations using estimations of the parameter  $p$  of the GSM for the different markers. As *MSVAR* only considers a strict SMM, *MIGRAINE* analyses are thus expected to give more accurate results. However, only three markers among the 14 used are dinucleotide loci, which likely explains the weak effect of mutation process misspecification on *MSVAR* analyses and thus the observed good agreement between *MIGRAINE* and *MSVAR* results.

Second, despite the fact that Goossens et al. (2005) showed a weak but significant population structure among sites, especially across the Kinabatangan River (e.g., between RS1 and RS2 pooled samples, see the Materials and Methods section), the strong similarity observed among all our results for site-specific and pooled samples suggests that this factor does not have a strong effect on the estimations. Sharma et al. (2012) similarly conclude of no major effect of population structure on their inferences.

## Conclusion

This work shows that our new inference method seems very competitive compared with alternative methods, such as *MSVAR*. However, our simulation tests also showed some 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.

Finally, 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 intralocus recombination, or single nucleotide polymorphisms (SNPs). The current version of *MIGRAINE* can consider three mutation models for allelic data (KAM, SMM, and GSM). It does not allow SNP data analysis, except under a KAM model with  $K=2$  allelic states. However, such a mutation model may not be adapted for SNP data because of the possibility of recurrent and backward mutations. For this reason, we did not test such analyses but a mutation model for SNPs is currently being implemented in *MIGRAINE*.

Given the relatively large computation times of our method, all analyses will clearly only be tractable for a limited number of markers (e.g.,  $<10,000$ ), 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 nonrecombining 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 historical inferences.

## Materials and Methods

### Simulation Study

A first set of simulations aims at testing the power of the algorithm to detect contractions and the accuracy of the parameters estimates when the duration ( $D$ ) or the strength of the contraction ( $\theta_{\text{anc}}$ ) varies. The mutation process considered is an SMM over a range of 200 alleles. These experiments are presented in table 9. We also reanalyzed the 60 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 supplementary table S2, Supplementary Material online, and the comparison results are presented in section B in the supplementary material, Supplementary Material online.

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 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, 2004; Estoup et al. 2001), and the largest, ever reported value (Fitzsimmons 1998; Peery et al. 2012). We have also added data sets drawn with the KAM to those simulations, which might be seen as a GSM with  $p = 1.0$ . A first set of analyses tests the robustness to

**Table 9.** Simulated Demographic Scenarios with an SMM.

Case	$D$ (T)	$\theta$ (N)	$\theta_{\text{anc}}$ ( $N_{\text{anc}}$ )
[0]	1.25 (200)	0.4 (200)	40.0 (20,000)
[1]	0.025 (10)	0.4 (200)	40.0 (20,000)
[2]	0.0625 (25)	0.4 (200)	40.0 (20,000)
[3]	0.125 (50)	0.4 (200)	40.0 (20,000)
[4]	0.25 (100)	0.4 (200)	40.0 (20,000)
[5]	0.5 (200)	0.4 (200)	40.0 (20,000)
[6]	2.5 (1,000)	0.4 (200)	40.0 (20,000)
[7]	3.5 (1,400)	0.4 (200)	40.0 (20,000)
[8]	5 (2,000)	0.4 (200)	40.0 (20,000)
[9]	7.5 (3,000)	0.4 (200)	40.0 (20,000)
[10]	1.25 (200)	0.4 (200)	2.0 (1,000)
[11]	1.25 (200)	0.4 (200)	4.0 (2,000)
[12]	1.25 (200)	0.4 (200)	8.0 (4,000)
[13]	1.25 (200)	0.4 (200)	12.0 (6,000)
[14]	1.25 (200)	0.4 (200)	24.0 (16,000)
[15]	1.25 (200)	0.4 (200)	120.0 (60,000)
[16]	1.25 (200)	0.4 (200)	400.0 (200,000)

misspecification of the mutation process. Indeed, we have simulated under a GSM but inferred under an 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  $p$ . Because mutation processes generally differ between loci, we also considered a situation with a combination of loci with different mutation models and different mutation rates: 80% of the loci are simulated under a GSM with  $p = 0.22$  and 20% under an SMM, and each locus has a mutation rate drawn from a Gamma distribution with parameters 2 and 0.0005, which gives a mean mutation rate of 0.001. For this situation with variable mutation processes, inference is also done using 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, described in table 3, were simulated under a GSM with  $p = 0.22$ . A first group of data sets simulates local within-population structure according to an IBD model. It thus aims at testing the robustness of the inferences to the assumption of panmixia by considering nonrandom 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. Under the island model of population structure, few additional simulations were run to test 1) the influence of the mutation model and 2) the effect of sampling a single gene per deme. Those simulations are presented in section C in the [supplementary material, Supplementary Material](#) online, and discussed in the main text.

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  $n_l = 10$  unlinked microsatellite loci, except for a few situations where we indicate that 25 or 50 instead of ten 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 material, Supplementary Material](#) online), have been produced with a new version of the `IBDSIM` software (Leblois et al. 2009) that considers continuous changes of population sizes.

## Validation

In all simulation experiments, the true (simulated) values of the parameters of interest are compared with the estimated values. The estimation bias and error, assessed by the relative mean bias and RRMSE, are reported, as well as the proportion of data sets for which a contraction or a false expansion signal is significantly detected (CDR and FEDR, respectively). Furthermore, the accuracy of the inference methodology is assessed by mean of profile LRTs (Cox and Hinkley 1974; Severini 2000). The coverage properties of the CI computed from the smoothed likelihood surface are tested through the ECDF 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.

## Orangutan Data Set Analysis

Finally, we analyzed an orangutan (*Pongo pygmaeus*) data set from Goossens et al. (2006), sampled in 2001 in the Lower Kinabatangan food plain in Eastern Sabah, Malaysia from feces and hairs, and genotyped at 14 microsatellite loci. Our method was first applied on each subsamples S1–S6, S8, and S9 as described in Goossens et al. (2005). Those samples correspond to different more or less isolated sampling sites on each side of the Kinabatangan River (see fig. 1 in Goossens et al. 2005). Subsample S7 was not analyzed because it only contains seven individuals, whereas all others contain between 16 and 33 individuals. Goossens et al. (2005) showed a weak differentiation level within each side of the river ( $0.01 \leq F_{ST} \leq 0.04$ , with a mean of 0.02) but a stronger one between both sides of the river ( $0.03 \leq F_{ST} \leq 0.09$  and a mean of 0.06). For this reason, and because our simulation study shows an important effect of population structure on the detection and characterization of past population size contractions, we also analyzed three larger data sets by pooling some subsamples together: 1) RS1 is the pool of subsamples coming from the south side of the river (i.e., S1+S3+S6+S8); 2) RS2 the pool of subsamples from the north of the river (S2+S4+S5+S9), and 3) RS1+RS2 is the total sample from both sides of the river.

To compare our results with those from Goossens et al. (2006) and Sharma et al. (2012) that used `MSVAR` on the same data sets, we need to convert our estimates of scaled parameters  $\theta$ ,  $D$ , and  $\theta_{anc}$  into canonical parameters (i.e.,  $N$ ,  $T$ , and  $N_{anc}$ ), because only values of canonical parameters are reported in these two publications. We did this conversion by considering a fixed mutation rate of  $5 \times 10^{-4}$  mutation per locus per generation, a value commonly considered as a realistic average value in the literature (Dib et al. 1996; Ellegren 2000; Sun et al. 2012). Furthermore, to express the timing of the contraction in years, we used a generation time of 25 years as in Sharma et al. (2012), a value in better agreement with long-term field studies than 8 years, the value considered in Goossens et al. (2006).

All sample sizes, results for all subsample analyses as well as details about the parametrization used for the inferences are given in section D in the [supplementary material, Supplementary Material](#) online. Results for the pooled data sets RS1, RS2, and RS1+RS2 are presented in the main text, in the Results section and illustrated in [supplementary figure S4, Supplementary Material](#) online.

The `MIGRAINE` software, with the implementation of the above described methods, can be downloaded from the web page <http://kimura.univ-montp2.fr/~rousset/Migraine.htm> (last accessed July 28, 2014).

## Supplementary Material

Supplementary material, figures S1–S65, and tables S1–S5 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

The authors are grateful to L. Chikhi, J.-M. Cornuet, A. Estoup, J.-M. Marin, and three anonymous reviewers 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@CORALFISH 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 GENOTOU bioinformatics platforms, and the computing grids of ISEM and CBGP labs.

## References

- Albrechtsen A, Sand Korneliusen T, Moltke I, van Overseem Hansen T, Nielsen FC, Nielsen R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genet Epidemiol.* 33:266–274.
- Beaumont M. 1999. Detecting population expansion and decline using microsatellites. *Genetics* 153:2013–2029.
- Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. *Proc Natl Acad Sci U S A.* 98:4563–4568.
- Bhargava A, Fuentes F. 2010. Mutational dynamics of microsatellites. *Mol Biotechnol.* 44:250–266.
- Bonebrake T, Christensen J, Boggs C, Ehrlich P. 2010. Population decline assessment, historical baselines, and conservation. *Conserv Lett.* 3: 371–378.
- Chikhi L, Sousa VC, Luisi P, Goossens B, Beaumont MA. 2010. The confounding effects of population structure, genetic diversity and the sampling scheme on the detection and quantification of population size changes. *Genetics* 186:983–995.
- Colautti RI, Manca M, Viljanen M, Ketelaars HAM, Bürgi H, Macisaac HJ, Heath DD. 2005. Invasion genetics of the Eurasian spiny waterflea: evidence for bottlenecks and gene flow using microsatellites. *Mol Ecol.* 14:1869–1879.
- Comps B, Gömöry D, Letouzey J, Thiébaud B, Petit RJ. 2001. Diverging trends between heterozygosity and allelic richness during postglacial colonization in the European beech. *Genetics* 157:389–397.
- Cornuet JM, Beaumont MA. 2007. A note on the accuracy of PAC-likelihood inference with microsatellite data. *Theor Popul Biol.* 71: 12–19.
- Cornuet JM, Luikart G. 1996. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. *Genetics* 144:2001–2014.
- Cornuet JM, Santos F, Beaumont MA, Robert CP, Marin JM, Balding DJ, Guillemaud T, Estoup A. 2008. Inferring population history with DIY ABC: a user-friendly approach to approximate Bayesian computation. *Bioinformatics* 24:2713–2719.
- Cox DR, Hinkley DV. 1974. Theoretical statistics. London: Chapman & Hall.
- Cressie NAC. 1993. Statistics for spatial data. New York: Wiley.
- de Iorio M, Griffiths RC. 2004a. Importance sampling on coalescent histories. *Adv Appl Probab.* 36:417–433.
- de Iorio M, Griffiths RC. 2004b. Importance sampling on coalescent histories. II. Subdivided population models. *Adv Appl Probab.* 36: 434–454.
- de Iorio M, Griffiths RC, Leblois R, Rousset F. 2005. Stepwise mutation likelihood computation by sequential importance sampling in subdivided population models. *Theor Popul Biol.* 68:41–53.
- Dib C, Fauré S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, et al. 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152–154.
- Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* 29: 1969–1973.
- Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH. 2009. Ancestral population genomics: the coalescent hidden Markov model approach. *Genetics* 183:259–274.
- Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference. *Trends Genet.* 16:551–558.
- Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 5:435–445.
- Emerson B, Paradis E, Thébaud C. 2001. Revealing the demographic histories of species using DNA sequences. *Trends Ecol Evol.* 16: 707–716.
- Estoup A, Wilson IJ, Sullivan C, Cornuet JM, Moritz C. 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*. *Genetics* 159:1671–1687.
- Faurby S, Pertoldi C. 2012. The consequences of the unlikely but critical assumption of stepwise mutation in the population genetic software, MSVAR. *Evol Ecol Res.* 14:859–879.
- Felsenstein J. 1992. Estimating effective population size from sample sequences—inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet Res.* 59:139–147.
- Fitzsimmons NN. 1998. Single paternity of clutches and sperm storage in the promiscuous green turtle (*Chelonia mydas*). *Mol Ecol.* 7:575–584.
- Frankham R, Lees K, Montgomery M, England P, Lowe E, Briscoe D. 2006. Do population size bottlenecks reduce evolutionary potential? *Anim Conserv* 2:255–260.
- Garza JC, Williamson EG. 2001. Detection of reduction in population size using data from microsatellite loci. *Mol Ecol.* 10:305–318.
- Girod C, Vitalis R, Leblois R, Fréville H. 2011. Inferring population decline and expansion from microsatellite data: a simulation-based evaluation of the Msvar method. *Genetics* 188:165–179.
- Gonser R, Donnelly P, Nicholson G, Di Rienzo A. 2000. Microsatellite mutations and inferences about human demography. *Genetics* 154: 1793–1807.
- Goossens B, Chikhi L, Ancrenaz M, Lackman-Ancrenaz I, Andau P, Bruford MW. 2006. Genetic signature of anthropogenic population collapse in orang-utans. *PLoS Biol.* 4:e25.
- Goossens B, Chikhi L, Jalil MF, Ancrenaz M, Lackman-Ancrenaz I, Mohamed M, Andau P, Bruford MW. 2005. Patterns of genetic diversity and migration in increasingly fragmented and declining orang-utan (*Pongo pygmaeus*) populations from Sabah, Malaysia. *Mol Ecol.* 14:441–456.
- Griffiths RC, Tavaré S. 1994a. Ancestral inference in population genetics. *Stat Sci.* 9:307–319.
- Griffiths RC, Tavaré S. 1994b. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.
- Guillot G, Leblois R, Coulon A, Frantz AC. 2009. Statistical methods in spatial genetics. *Mol Ecol.* 18:4734–4756.
- Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, Gregersen P, Pe'er I. 2012. The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol.* 29:473–486.
- Heller R, Chikhi L, Siegmund HR. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8:e62992.
- Hey J. 2010. Isolation with migration models for more than two populations. *Mol Biol Evol.* 27:905–920.
- Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167:747–760.
- Hey J, Nielsen R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc Natl Acad Sci U S A.* 104:2785–2790.
- Keller LF, Waller DM. 2002. Inbreeding effects in wild populations. *Trends Ecol Evol.* 17:230–241.
- Kuhner MK. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. *Bioinformatics* 22:768–770.

- Lande R. 1988. Genetics and demography in biological conservation. *Science* 241:1455–1460.
- Lawton-Rauh A. 2008. Demographic processes shaping genetic variation. *Curr Opin Plant Biol.* 11:103–109.
- Leblois R, Estoup A, Rousset F. 2009. IBDSim: a computer program to simulate genotypic data under isolation by distance. *Mol Ecol Resour.* 9:107–109.
- Leblois R, Estoup A, Streiff R. 2006. Habitat contraction and reduction in population size: does isolation by distance matter? *Mol Ecol.* 15: 3601–3615.
- Mailund T, Halager AE, Westergaard M, Dutheil JY, Munch K, Andersen LN, Lunter G, Prüfer K, Scally A, Hobolth A, et al. 2012. A new isolation with migration model along complete genomes infers very different divergence processes among closely related Great Ape species. *PLoS Genet.* 8:e1003125.
- Marjoram P, Tavaré S. 2006. Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet.* 7: 759–770.
- Meuwissen TH, Goddard ME. 2007. Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* 176:2551–2560.
- Nielsen R, Beaumont MA. 2009. Statistical inferences in phylogeography. *Mol Ecol.* 18:1034–1047.
- Ohta T, Kimura M. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genet Res.* 22:201–204.
- Palamara PF, Lencz T, Darvasi A, Pe'er I. 2012. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet.* 91:809–822.
- Peery MZ, Kirby R, Reid BN, Stoelting R, Doucet-Bër E, Robinson S, Vázquez-Carrillo C, Pauli JN, Palsbøll PJ. 2012. Reliability of genetic bottleneck tests for detecting recent population declines. *Mol Ecol.* 21:3403–3418.
- Peter B, Wegmann D, Excoffier L. 2010. Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure. *Mol Ecol.* 19:4648–4660.
- Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. 1999. Population growth of human Y chromosome microsatellites. *Mol Biol Evol.* 16:1791–1798.
- Ptak SE, Przeworski M. 2002. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends Genet.* 18: 559–563.
- Reusch TBH, Wood TE. 2007. Molecular ecology of global change. *Mol Ecol.* 16:3973–3992.
- Rousset F, Leblois R. 2007. Likelihood and approximate likelihood analyses of genetic structure in a linear habitat: performance and robustness to model mis-specification. *Mol Biol Evol.* 24: 2730–2745.
- Rousset F, Leblois R. 2012. Likelihood-based inferences under a coalescent model of isolation by distance: two-dimensional habitats and confidence intervals. *Mol Biol Evol.* 29:957–973.
- Schneider S, Excoffier L. 1999. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 152:1079–1089.
- Schwartz M, Luikart G, Waples R. 2007. Genetic monitoring as a promising tool for conservation and management. *Trends Ecol Evol.* 22: 25–33.
- Severini TA. 2000. Likelihood methods in statistics. Oxford: Oxford University Press.
- Sharma R, Arora N, Goossens B, Nater A, Morf N, Salmona J, Bruford MW, Van Schaik CP, Krützen M, Chikhi L. 2012. Effective population size dynamics and the demographic collapse of Bornean orangutans. *PLoS One* 7:e49429.
- Spencer CC, Neigel JE, Leberg PL. 2000. Experimental evaluation of the usefulness of microsatellite DNA for detecting demographic bottlenecks. *Mol Ecol.* 9:1517–1528.
- Stephens M, Donnelly P. 2000. Inference in molecular population genetics (with discussion). *J R Stat Soc Series B Stat Methodol.* 62: 605–655.
- Storz J, Beaumont M. 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* 56: 154–166.
- Sun J, Helgason A, Masson G, Ebenesersdottir S, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, et al. 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet.* 44: 1161–1165.
- Theunert C, Tang K, Lachmann M, Hu S, Stoneking M. 2012. Inferring the history of population size change from genome-wide SNP data. *Mol Biol Evol.* 29:3653–3667.
- Wakeley J. 1999. Nonequilibrium migration in human evolution. *Genetics* 153:1863–1871.
- Williams B, Nichols J, Conroy M. 2002. Analysis and management of animal populations: modeling, estimation, and decision making. San Diego (CA): Academic Press.
- Wright S. 1951. The genetical structure of populations. *Ann Eugen.* 15: 323–354.