Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12347

Theoretical expectations of the Isolation–Migration model of population evolution for inferring demographic parameters

Maud C. Quinzin[1*]†, Francois Mayer[1,2], Nora Elvinger[1,3] and Patrick Mardulyn[1]

[1]*Evolutionary Biology and Ecology, Université Libre de Bruxelles (ULB), Avenue F.D. Roosevelt 50, 1050 Brussels, Belgium;* [2]*Biological Control and Spatial Ecology, Université Libre de Bruxelles (ULB), Avenue F.D. Roosevelt 50, 1050 Brussels, Belgium;* [3]*Natural History Museum Luxembourg, Luxembourg, Luxembourg*

Summary

1. The Bayesian inference of demographic parameters under an Isolation–Migration (IM) model of population evolution offers a major improvement over previously available approaches. This method is implemented in a popular program, IMa, widely used in population genetic studies.

2. While the robustness of the method to deviations of the IM model has previously been evaluated, we assess the performance of the program with two populations when the model used to generate the analysed data meets the assumptions of the IM model completely; the goal is to identify the conditions under which the method works best. Overall, we test eighteen sets of conditions and analyse ± 500 simulated data sets, for a total of over 200,000 hours of analyses using a large computer cluster.

3. Although we find clear differences in quality estimates among models, the best ranges of demographic parameter values to infer accurate estimates differ among parameters. Divergence time is best estimated in the absence of gene flow and when population sizes are large compared to divergence time. In contrast, the classic population parameter ϑ ($= 4N\mu$) is best estimated, for the two current populations, when divergence time is large compared to population size, with or without migration. The parameter is always poorly estimated in the case of the ancestral population. While it is possible to distinguish between scenarios with or without gene flow, estimating the extent of gene flow, when different from 0, is associated with relatively high error rates. In general, increasing the number of loci or the sample size reduces the variance and credible interval of the estimates, and only for the migration rate, it slightly improves the accuracy of the estimate as well. Increasing the prior distribution range of a parameter can dramatically increase that of its posterior distribution. Surprisingly, differences are highlighted among the estimates inferred from sequences generated by different simulation programs, especially for the simulation program SIMDIV.

4. Overall, the performances of the method shown here probably reflect the limitation of the method in general and/or of the historical information contained in DNA sequence data.

Key-words: Bayesian inference, coalescence simulations, divergence time, effective population size, migration rate, performance assessment

Introduction

Early population genetic studies have often characterized differentiation among populations, using molecular markers, as an indirect mean to estimate gene flow (e.g. Slatkin 1985, 1987; Neigel 1997). For most natural systems, however, this approach is flawed because a given level of population differentiation may correspond to different levels of gene flow, depending on the history shared by these populations. For example, a low differentiation between two populations may reflect high gene flow among them, but also a recent common history without current migration. Only if populations have reached a state of migration–drift equilibrium (i.e. when the level of population differentiation reflects the extent of these two antagonistic forces), which is probably seldom the case in natural systems, can a direct relationship between population differentiation and gene flow be established (Slatkin 1985; Bohonak 1999; Whitlock & McCauley 1999).

The development, by Nielsen & Wakeley (2001) and Hey & Nielsen (2004), of a Markov chain Monte Carlo (MCMC) approach (in a likelihood or Bayesian framework) for the inference of standard demographic parameters (population size, migration rates, divergence times) under a more realistic 'Isolation–Migration' (IM) model of population

*Correspondence author. E-mail: maudquinzin@gmail.com

†Present address: OD Taxonomy and Phylogeny, RBINS – Royal Belgian Institute of Natural Sciences, 1000 Brussels, Belgium

evolution offered a major improvement to the estimation of gene flow and other demographic parameters in population studies. Its most attractive feature lies in its possibility to estimate demographic parameters in non-equilibrium situations (i.e. when effects of genetic drift on population differentiation are either stronger or weaker than those of migration). Implemented in two popular programs, IM and IMa(2) (Hey & Nielsen 2004, 2007), this approach has been widely adopted for evolutionary inference, as evidenced by its use in numerous studies during the last decade (e.g. Kotlík *et al.* 2006; Nunes *et al.* 2010; Prada & Hellberg 2013). Their intended primary use was to estimate divergence time and migration rates simultaneously, in the hope of being able to distinguish between scenarios of population divergence with and without gene flow (Nielsen & Wakeley 2001).

It is in general highly desirable to evaluate the performances of any new method of historical inference before applying it to real data; however, this task is inherently challenging, because we have no access to natural biological systems for which the historical parameters are known a priori (with a sufficiently high level of certainty), especially when studying evolutionary events spanning thousands of years or more. As an alternative for such purpose, the computer simulation of molecular markers (e.g. DNA sequences) according to pre-defined models of population evolution, including the specific values of each parameters, offers yet a powerful tool; it has been widely used in the past to identify conditions under which a method exhibits sufficient statistical power and/or is subject to various types of bias (e.g. Abdo, Crandall & Joyce 2004; Faubet, Waples & Gaggiotti 2007).

Although the performances of the IM and IMa computer programs have already been evaluated using computer simulations (Becquet & Przeworski 2009; Strasburg & Rieseberg 2010, 2011), such evaluation studies have focused primarily on situations that deviate from the IM model, for the purpose of testing the robustness of that method of inference. Here, we attempt to complement the evaluation by exploring cases in which the data have been generated within the boundaries of the IM model. While limiting our investigation to two population models (more than two populations can now be modelled; Hey 2010) like these previous studies did, we explored a wider range of parameter values within the boundaries of the IM model. More specifically, our goals were (i) to identify the conditions under which the IMa2 software works best, and (ii) to evaluate how accurate the estimations provided by this method are, when all the assumptions of the IM model are met. Since, by definition, real world data never entirely conform to the assumptions of a model, the quality of the estimates derived in this work can be considered as an upper limit to the quality that can be expected with empirical DNA sequence data. In this sense, this study explores the theoretical expectations of the IM model for inferring demographic parameters. It was achieved thanks to the extensive use of a large computer cluster, totalling more than 200 000 h of analyses with IMa2.

Methods

SIMULATIONS

All simulations of DNA sequence data were restricted to the basic IM model of two populations, defined by six parameters, each scaled by the mutation rate μ: population divergence time ($t = T \times \mu$), migration rates between daughter populations ($m = m/\mu$), and ϑ (4Nμ, with 2N = population size) for each of the daughter populations and the ancestral population. By varying ϑ and the time of divergence t, we generated four different demographic models, each with or without gene flow (i.e. 8 different versions of the IM model), spanning the spectrum of realistic conditions within the IM model framework (Fig. 1). The choice of parameter values was guided by those found in Becquet & Przeworski (2009) and Strasburg & Rieseberg (2010, 2011), and close to those found in many real data sets. We generated data sets to be analysed with IMa2 by simulating evolution of DNA sequences along the eight versions of the model. In an initial step, we used and compared three commonly used coalescence simulation programs: make sample (MS; Hudson 2002) in conjunction with Seq-Gen (Rambaut & Grassly 1997) simply referred as MS in this study, Simcoal 2 (Laval & Excoffier 2004) and SIMDIV (Hey 2010; Wang & Hey 2010). This comparison was intended to verify the absence of bias in the simulation programs. For this purpose, samples of 10 and 40 sequences per population, for five independent loci, each with an identical mutation rate, were simulated under Model 1.

Because it would have been impractical to perform all subsequent tests using sequences generated by the three simulation programs, we chose to limit all other analyses to a single simulation program. SIMDIV was chosen because it was developed by the same research group that created IMa2, and therefore, both programs presumably rely on the same model of coalescence with identical parameter definition. This is in line with our purpose of testing, the IMa2 estimation program for the case in which the coalescent model used for the estimation analyses meets the IM model assumptions. For all following simulations, the software SIMDIV was thus used, along with samples of 40 sequences per population, except for those analyses that specifically tested the effect of sample size on the accuracy of the estimations. For comparing among the four tested demographic models, each with or without gene flow, we generated data sets of five loci, each with an identical mutation rate. Finally, Model 3 without gene flow was chosen for additional simulations in order to evaluate the impact of (i) the number of loci (5, 10, 20 loci), (ii) the sample sizes (10, 40, 100 sequences per population), and (iii) the upper bounds of the prior distributions (see below) on the parameter estimates. We chose Model 3 because we obtained the best estimations under its associated demographic conditions, during the previous analysis steps.

For each set of parameter values tested, we simulated at least ten data sets for analysis with IMa2, but most of the times twenty to fifty sets were simulated and analysed. The Hasegawa-Kishino-Yano (HKY) model of nucleotide substitution (Hasegawa, Kishino & Yano 1985) was used in all simulations, and, for convenience, all loci shared the same mutation rate. Note that the base model used by Strasburg & Rieseberg (2010, 2011), from which different types of deviations were derived, included a parameter ϑ four times larger than the divergence time t, which makes it similar to the situation of Model 3 in this study. In Becquet & Przeworski (2009), t was twice as large as ϑ, which resulted in an intermediate situation between our Models 1 and 2.

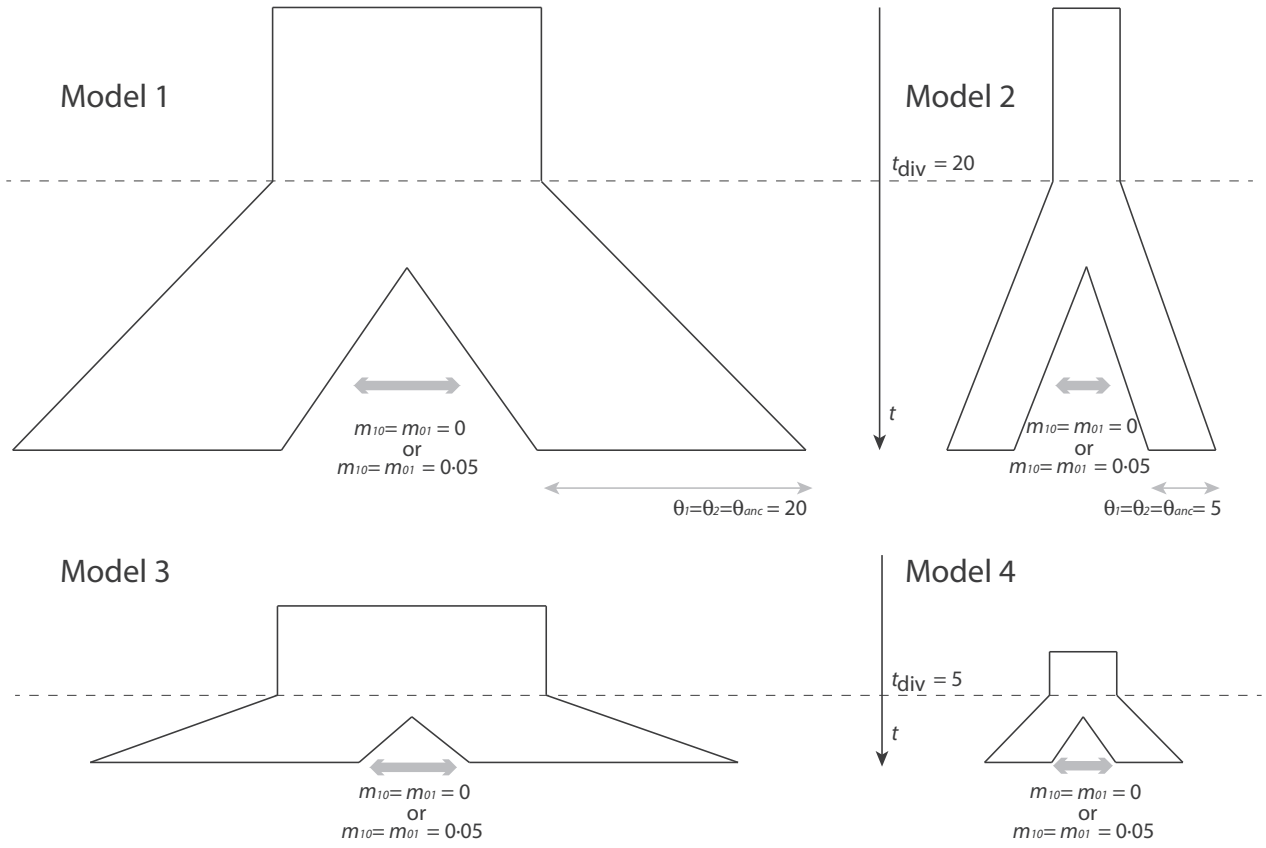

Fig. 1. Four models, each implemented in the presence or absence of gene flow, used here to simulate DNA sequences for the purpose of evaluating the inference of demographic parameters under an Isolation–Migration model. All model parameters are standardized by the mutation rate μ: the standardized migration rate $m = m/\mu$, the standardized divergence time $t = \mu t$, and the classic population parameter $\vartheta = 4\ N\ \mu$, where $2N$ is the population size (number of gene copies).

IMA2 ANALYSES

All analyses were performed with the software IMa2, version 12.17.09 or 10.13.10 (https://bio.cst.temple.edu/~hey/software/software.htm#IM). To ensure convergence among MCMC runs, each data set was analysed three times under the exact same conditions but with different random number seeds. We then checked whether we obtained similar parameter estimates among runs. Each run included between 40 and 120 MCMC chains with geometric heating and lasted at least 2 million steps after a burn-in of 100 000 steps. As recommended in the manual, the number of MCMC chains per run or/and the number of steps per chain was/were raised until all analyses resulted in effective sample size (ESS) values for all parameters above 50 (with ESS values >10 000 in most cases), indicating a good mixing of MCMC chains and good coverage of genealogy/parameter space. In most analyses, maximum values for the prior ranges were set to $t_{max} = 100$, $\vartheta_{max} = 100$, and $m_{max} = 1$. For evaluating the influence of the prior distribution on the estimations, analyses were also run with these maximum values divided by 2 or multiplied by 2·5. The nucleotide substitution model was set to HKY, the one used for all simulations.

STATISTICAL COMPARISONS

For comparing the mean estimates inferred with the simulated data from the three simulation programs tested, we first determined whether the estimated values conformed to a normal distribution (Shapiro and Wilk normality test) and that the variances around the means of the estimates were homogeneously distributed (homoscedasticity; Bartlett's test). When one or both conditions were not met, we applied a square-root transformation of the data and conducted the same two tests on the transformed values. If both conditions were met (either on the original values or on the square-root transformations), we conducted a one-way ANOVA, followed by a Tukey's HSD (Honest Significant Difference) test for pairwise comparison of simulation programs. Otherwise, we applied a Kruskal–Wallis nonparametric test (analysis of variance by ranks), followed by the multiple comparison test 'kruskalmc' (implemented in the R-package *pgirmess*), with a significance level of 0·05. All statistical analyses were conducted in the R environment (R Core Team 2014).

Results

The three MCMC runs conducted with IMa2 on each data set resulted each time in very similar estimates, indicating strong convergence among runs. Additional runs were therefore never needed in our analyses. Point estimates, along with their estimated 95% highest posterior density interval (hereafter referred to as credible interval; CI), are graphically displayed in Figs 2–4, or summarized in Tables 2 and 3 (for results graphically displayed in Figs S3 and S4). While we obtained estimates for up to fifty IMa2 analyses (each for a different simulated data set) per set of conditions tested, we present only ten estimates per set on each graph for clarity; it appears sufficient

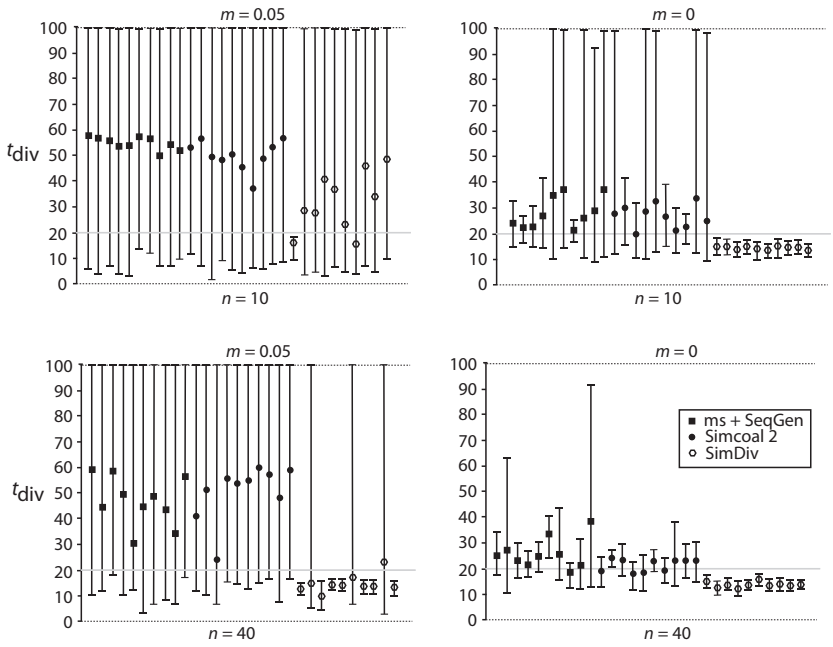

Fig. 2. Divergence time estimates obtained from DNA sequences simulated under Model 1, with and without migration, with three different simulation programs. Point estimates are presented along with their 95% highest posterior density interval (referred to as the *credible interval* in the text). A grey line indicates the true value of the parameter (i.e., the value implemented in the model for simulating the sequences), and dashed lines delimit the range of the prior distribution for that parameter. Estimates are provided for two different sample sizes (n = 10 and 40 per population).

to characterize the variation among estimates. However, full results are detailed in Supporting information (Tables S2–S6).

COMPARING SIMULATION PROGRAMS

We compared parameter estimates obtained from sequence data sets generated with all three simulation programs (Figs 2, S1 and S2 and Table S1). Divergence time estimates are clearly better when no migration has occurred between the two populations, and increasing the number of sampled sequences (from 10 to 40) improves them in many cases (Fig. 2). When migration does occur and the number of sequences per population is restricted to 10, most analyses result in extremely large CIs, spanning almost the entire prior range, and the position of the estimate is often simply located in the middle of this interval, suggesting there is too little historical signal within the data to gain any valuable information. For ϑ parameters, the estimates for the ancestral population are non-informative (large CIs) but are reasonably good for the two daughter populations. The examination of estimates of migration rates reveals a general trend of overestimation.

Surprisingly, aside from the general trends discussed above, estimates from data sets generated with different simulation programs appear quite different. The difference between estimates obtained from sequences generated with SIMDIV and those generated with the two other computer programs is in most cases large and statistically significant (P-value <0·05), while no significant difference is observed between MS and Simcoal 2, (Table S1, S7 and S8). These differences are observed with sample sizes of 10 and 40, and with or without gene flow between populations. Estimates of divergence time obtained from SIMDIV sequences display much shorter CIs, but are often biased towards values lower than the true value (grey line on the graphs), which is not included in the CI (Fig. 2). For ϑ, estimates for the daughter populations are reasonably close to the true value when obtained from sequences generated with Simcoal 2 and MS, but are largely underestimated when obtained from sequences generated with SIMDIV (Fig. S1); ϑ estimates for the ancestral population (q_{anc}), however, are always poor, whatever program was used to simulate the sequences (Fig. S1). Conversely, migration rates are most often similarly estimated with sequences generated by the three simulation programs, and we found improved estimates for larger data sets (n = 40) in all cases (Table 1; Fig. S2). In general, IMa2 often underestimates parameter values when used on SIMDIV data sets (in more than 60% of the cases; otherwise true value included in credible interval), most clearly for two parameters, divergence time and ϑ for the ancestral population, when there is no migration between them (n = 40 and n = 10), or in the presence of gene flow (when n = 40).

COMPARING PARAMETER ESTIMATES AMONG MODELS

Comparison of estimates for the four parameters (t_{div}, ϑ, ϑ_{anc}, and m) across demographic models reveals strong variation among models (Table 2). To evaluate these estimates, we focus on (i) the distance between them and the true value (i.e. absolute error), (ii) the inclusion of the latter in the CIs, and (iii) the range of the CIs. We try hereafter to define the best demographic conditions for the estimations, separately for each parameter (Table 1).

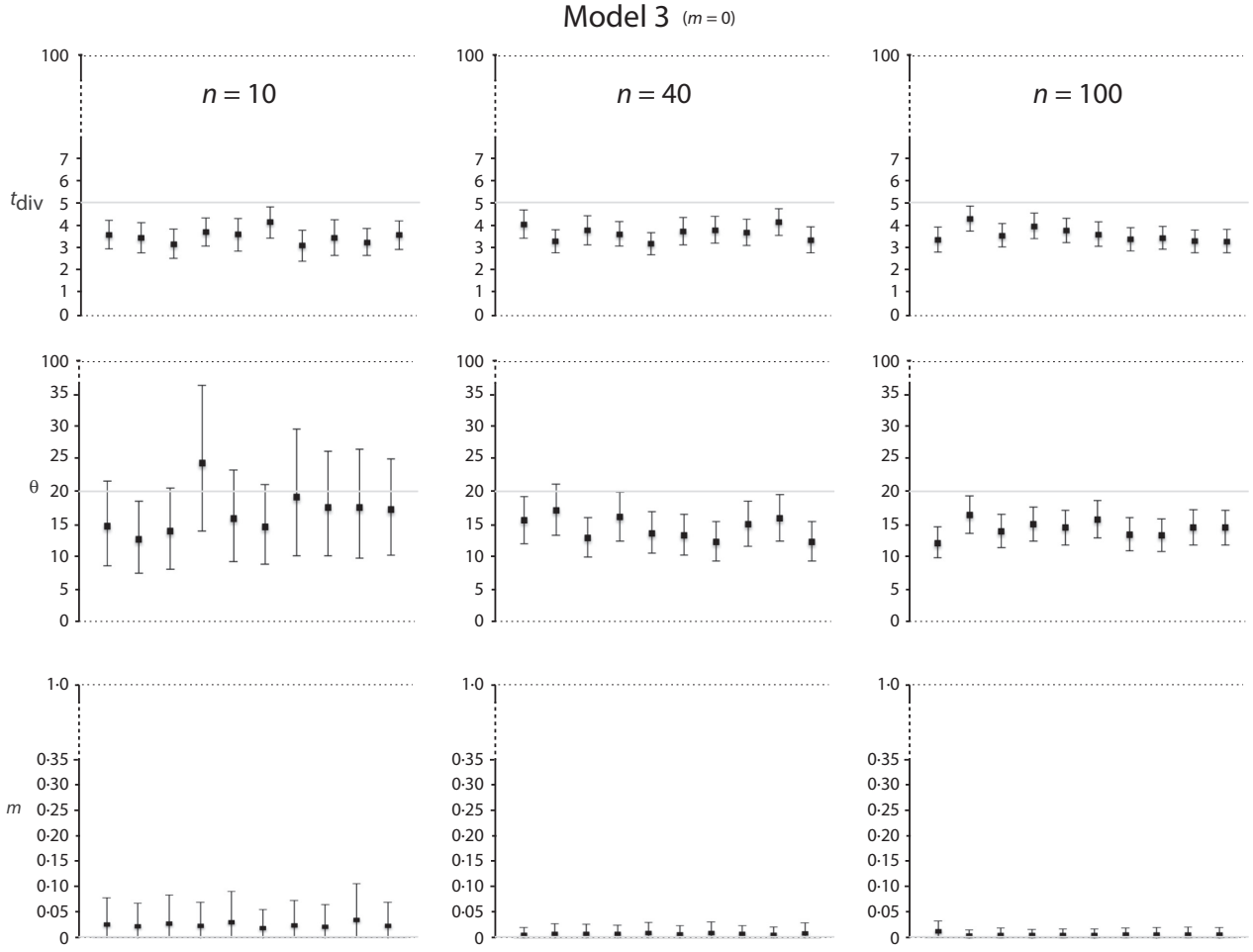

Fig. 3. Estimates of three demographic parameters obtained by analysing samples of 10, 40 or 100 DNA sequences per population, generated by simulations along Model 3 without migration. Point estimates are presented along with their 95% highest posterior density interval (referred to as the *credible interval* or CI in the text). A grey line indicates the true value of the parameter, and dashed lines delimit the range of the prior distribution for that parameter.

As already hinted in the previous section (comparing simulation programs with Model 1), divergence time is best estimated when no migration occurs between the two populations (Fig. 2 and Table 2). With inferences from sequences generated under Model 2, relatively good estimates are obtained in most cases in the absence of migration, but are associated with the largest CIs (almost identical to the prior ranges). In general, point estimates appear closest to the true value with Model 3. For Models 1, 3 and 4, estimates are associated with small CIs that, however, never include the true value. In the presence of migration, IMa2 poorly infers divergence time on sequences generated under Model 3 (estimates far from the true value and associated with very large CIs) and better performs on sequences generated under Model 1 (estimates often close to the true value and associated with short CIs). Relatively shorter CIs are also associated with Model 2, although point estimates are systematically much smaller than the true value in that case. Overall, divergence times are best estimated under conditions of Model 1 in the presence of gene flow.

Focusing on migration rate m (Table 2, and Fig. S3B), estimates are worse in the presence of gene flow, as they are generally more distant from the true value and associated with much larger CIs. Sequences generated under Models 3 and 1 appear to provide better estimates in this case. When migration is implemented in the model, much more variation is observed among estimates of m. In that case, Model 1 seems to offer the best data sets for estimation in terms of approaching the true value, of providing shorter CIs, and of generating consistent results. On the contrary, Model 4 provided the worst estimates on all these accounts. It is worth noting that all estimates from sequences generated under demographic models that did not implement migration ($m = 0$) included the true value of zero in the associated CI, while most estimates from sequences generated under models including migration ($m = 0{\cdot}05$) did not include zero in their associated CI. For sequences generated under a model implementing migration, the true value is almost always included in the credible interval of the estimates, although the range of this interval can be quite large.

For parameter ϑ, sequences simulated under Model 2 ($m = 0$ or $0{\cdot}05$) and under Model 4 without migration

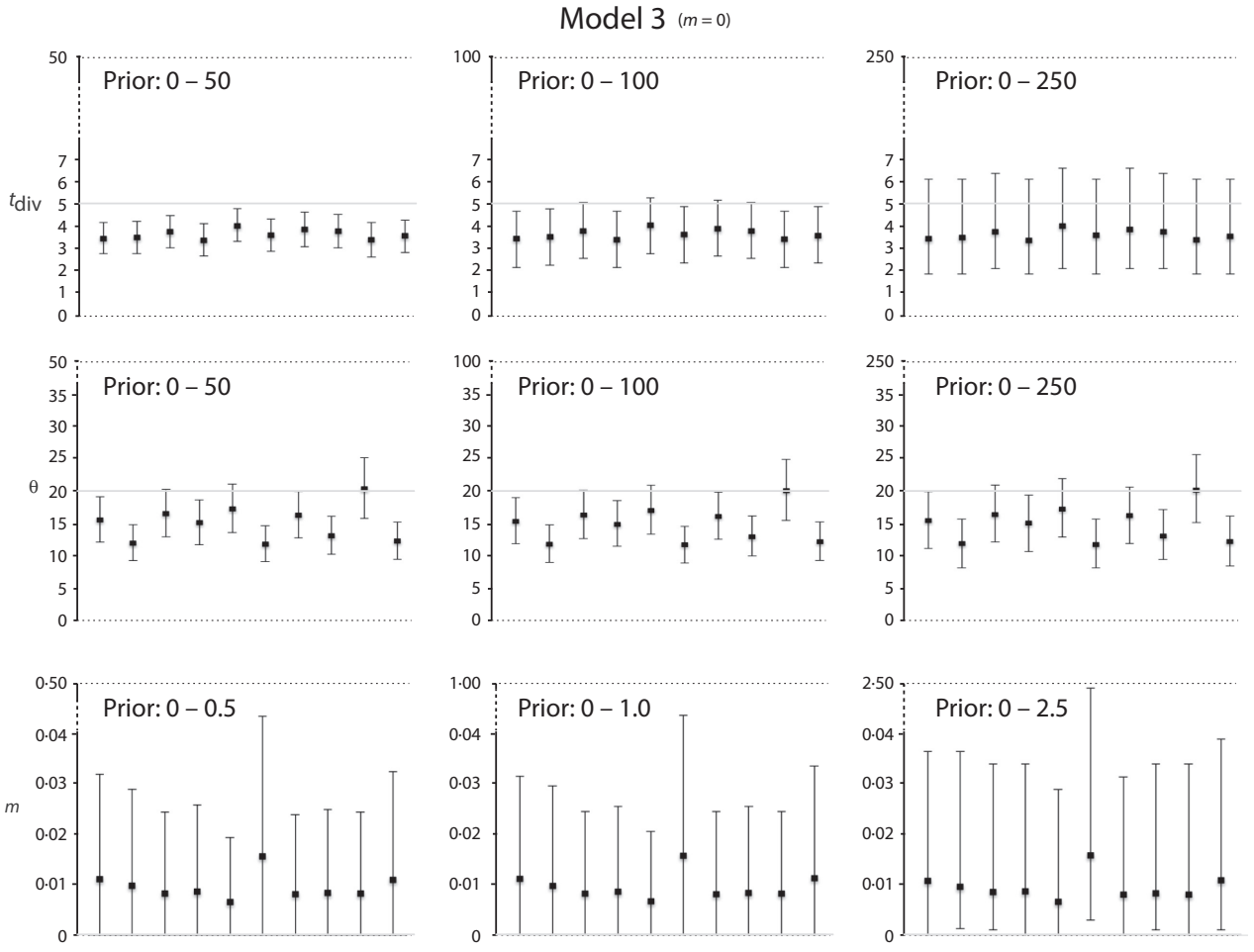

Fig. 4. Estimates of three demographic parameters obtained by analysing DNA sequences generated by simulations along Model 3 without migration, using three different prior distribution ranges. Point estimates are presented along with their 95% highest posterior density interval (referred to as the *credible interval* or CI in the text). A grey line indicates the true value of the parameter, and dashed lines delimit the range of the prior distribution for that parameter.

($m = 0$) seem to provide the best estimates (point estimate closest to true value; Table 2 and Fig. S3C). Here, differences in estimates between models with and without migration are much less obvious than with the two other parameters (t_{div} and m). Note that, for all models, the true value is systematically underestimated (only two exceptions out of eighty estimates). Also, in the cases with migration, Model 1 and 4 (in which ϑ and t_{div} are relatively similar) offer conditions under which the estimated CI is larger but includes more often the true value. Finally, the parameter ϑ for the ancestral population (ϑ_{anc}) is generally poorly estimated compared to that of the daughter populations,

Table 1. summary of best demographic conditions to estimate four demographic parameters, taking into account the position of point estimate and credible interval (CI) relative to true values, and size of CI

	Models with Migration	Models without migration
t_{div}	Estimates: Model 1 > 2 > 4 > 3 CI: Model 2 (often include true value) and 1 (small intervals never include true value), 4 and 3 (maximum intervals)	Estimates: Model 3 > 4, 2 > 1 CI: Model 3 > 4,1 (small intervals never include true value) > Model 2 (very large intervals, always include true value)
m	Estimates: Model 1 > 3, 2 > 4 CI: Model 1 > 3, 2 > 4	Estimates: Model 3, 1 > 4, 2 CI: Model 3, 1 > 2, 4
ϑ	Estimates: Model 2, 4 > 1 > 3 CI: Model 2, 4 (larger ranges but includes true value more often) > Model 1, 3	Estimates: Model 2, 4 > 1, 3 CI: Model 2, 4 > 1, 3
ϑ_{anc}	Estimates: Model 2, 3 > 1 > 4 CI: Model 3 (includes 0 less often), 2 > 1 > 4	Estimates: Model 2, 4 > 1 > 3 CI: Model 2 > 4 > 1 > 3

Divergence time, t_{div}; migration rate, m; daughter population effective sizes, ϑ; ancestral population effective size, ϑ_{anc}; >: generally associated with better results than.

Table 2. Quality of parameter estimates under the four different demographic models

$m = 0$		Model 1 $\vartheta = 20$, $t_{div} = 20$	Model 2 $\vartheta = 5$, $t_{div} = 20$	Model 3 $\vartheta = 20$, $t_{div} = 5$	Model 4 $\vartheta = 5$, $t_{div} = 5$
t_{div}	Mean error ± SD	6·31 ± 1·04	10·75 ± 7·24	1·30 ± 0·32	1·68 ± 0·29
	Mean CI range ± SD	5·18 ± 0·72	95·22 ± 0·88	1·16 ± 0·10	2·22 ± 0·33
	True value in CI	0	1	0	0
ϑ	Mean error ± SD	5·05 ± 1·27	1·16 ± 0·60	5·53 ± 1·76	1·09 ± 0·56
	Mean CI range ± SD	6·47 ± 0·35	2·32 ± 0·23	6·81 ± 0·68	2·45 ± 0·22
	True value in CI	0·2	0·5	0·2	0·5
ϑ_{anc}	Mean error ± SD	16·38 ± 1·46	4·82 ± 1·78	19·68 ± 0·08	3·34 ± 0·55
	Mean CI range ± SD	9·39 ± 3·09	22·02 ± 1·19	1·29 ± 0·13	4·45 ± 1·28
	True value in CI	0	1	0	0·2
m	Mean error ± SD	0·009 ± 0·0013	0·0155 ± 0·004	0·009 ± 0·0013	0·0204 ± 0·0019
	Mean CI range ± SD	0·0272 ± 0·0037	0·0462 ± 0·0109	0·0272 ± 0·0037	0·0605 ± 0·0057
	True value in CI	1	1	1	1
$m = 0·5$		Model 1 $\vartheta = 20$, $t_{div} = 20$	Model 2 $\vartheta = 5$, $t_{div} = 20$	Model 3 $\vartheta = 20$, $t_{div} = 5$	Model 4 $\vartheta = 5$, $t_{div} = 5$
t_{div}	Mean error ± SD	5·72 ± 2·04	30·81 ± 1·96	1·07 ± 1·35	20·80 ± 17·51
	Mean CI range ± SD	33·17 ± 42·97	94·79 ± 2·71	15·39 ± 9·40	89·03 ± 30·19
	True value in CI	0·3	0·6	0·6	0·9
ϑ	Mean error ± SD	3·45 ± 1·80	1·48 ± 0·45	5·15 ± 1·44	1·43 ± 0·73
	Mean CI range ± SD	7·66 ± 0·93	2·22 ± 0·20	7·41 ± 0·52	3·06 ± 0·38
	True value in CI	0·6	0·1	0·3	0·7
ϑ_{anc}	Mean error ± SD	13·71 ± 4·61	7·36 ± 0·25	5·41 ± 2·71	19·03 ± 16·46
	Mean CI range ± SD	21·63 ± 23·19	23·66 ± 0·16	23·77 ± 15·53	72·08 ± 29·72
	True value in CI	0·3	1	0·7	0·9
m	Mean error ± SD	0·0205 ± 0·0179	0·0484 ± 0·0419	0·045 ± 0·0395	0·0695 ± 0·0748
	Mean CI range ± SD	0·1023 ± 0·0243	0·1679 ± 0·0509	0·1558 ± 0·0544	0·2007 ± 0·1191
	True value in CI	1	1	1	0·9

Parameters: t_{div} = divergence time, ϑ = population parameter, ϑ_{anc} = ancestral population parameter, m = migration rate; mean error = average difference between true value and estimated value, SD = standard deviation, mean CI range = average range of credible interval, true value in CI = proportion of estimates for which the CI includes the true value.

although Model 2 ($m = 0$ or 0·05) and Model 3 (m = 0·05) seem to offer the best conditions for its estimation.

IMPACT OF THE AMOUNT OF DATA ON THE ESTIMATION

Number of loci

We investigated how much estimates could be improved by analysing an increasing number of loci (5, 10, 20; Table 3, Fig. S4). Overall, slightly shorter intervals, and lower variation among estimates, are obtained by increasing the number of loci to 10, then to 20. Surprisingly, the bias observed for t_{div} and ϑ (estimates lower than the true value) remains when analysing more loci. In contrast, the accuracy of the migration rate estimate increases when analysing data sets including more loci, both by reducing the length of the CIs and by approaching the true value.

Sample size

We also studied the effect of increasing the number of sampled sequences from 10 to 40 and 100 on parameter estimation (Fig. 3). No improvement was detected in the estimation of t_{div} when increasing the number of sampled sequences, probably because the estimation in these conditions (Model 3, $m = 0$) is already satisfactory with 10 sequences per population. However, we already demonstrated how increasing the number of sequences can improve the estimate of t_{div} if the initial estimation (i.e. from only 10 sampled sequences per population) is poor, as evidenced by a large CI (Model 1, Fig. 2). Like increasing the number of loci, increasing the number of sequences increases the accuracy of point estimates of m and reduces the CIs around the estimates for both parameters ϑ and m. However, the number of ϑ estimates for which the CI includes the true value decreases, reaching zero with $n = 100$.

IMPACT OF PRIOR DISTRIBUTION ON THE ESTIMATION

Finally, we verified if the size of the prior range of a parameter may influence its estimation. In general, the prior distribution had little effect on the estimate itself. However, in some cases, we observed that the size of the prior range and that of the CI are correlated (Fig. 4; see also Figs S5–S8 for typical posterior distributions). This was most striking with parameter t_{div}, where making the prior range wider has clearly increased the size of the credible interval. While the interval never included the true value with the shortest tested prior, it always did with the largest one. A similar trend, much less pronounced, is

Table 3. Quality of parameter estimates when increasing the number of loci (n = 5, 10, 20), under Model 3 (ϑ = 20, t_{div} = 5) without migration (m = 0)

		5 loci	10 loci	20 loci
t_{div}	Mean error ± SD	1·30 ± 0·32	1·36 ± 0·20	1·36 ± 0·14
	Mean CI range ± SD	1·16 ± 0·10	0·88 ± 0·03	0·72 ± 0·01
	True value in CI	0	0	0
ϑ	Mean error ± SD	5·53 ± 1·76	5·42 ± 1·44	4·91 ± 0·56
	Mean CI range ± SD	6·81 ± 0·68	5·06 ± 0·36	3·91 ± 0·10
	True value in CI	0·1	0·1	0
m	Mean error ± SD	0·0090 ± 0·0013	0·0044 ± 0·0003	0·0025 ± 0·0005
	Mean CI range ± SD	0·0272 ± 0·0365	0·0152 ± 0·0007	0·0118 ± 0·0004
	True value in CI	1·0	1·0	1·0

Parameters: t_{div} = divergence time, ϑ = population parameter, m = migration rate; mean error = average difference between true value and estimated value, SD = standard deviation, mean CI range = average range of credible interval, true value in CI = proportion of estimates for which the CI includes the true value.

observed for the ϑ parameter. The most negative influence of an increased prior range is observed for the migration parameter m, where not only the size of the CI increased but also the CI less often included the true value.

Discussion

The IM model represents a major improvement for investigating demographic hypotheses in the presence of gene flow. Attempting to identify the best conditions under which we can infer demographic parameters with DNA sequences under the IM model framework, we cannot point to a single best historical scenario. Rather, differences in estimate quality among models varied largely among demographic parameters (divergence time, population sizes, gene flow) and among data sets (various sampling size, number of loci).

COMPARING SIMULATION PROGRAMS

One finding highlighted in the course of our investigation was the major differences observed among the outputs of the three tested simulation programs. Data generated by SIMDIV appeared significantly different from those produced by the two other programs, Simcoal and ms. This suggests a possible error in SIMDIV's code, which could explain the smaller CIs and the more systematically biased estimates (towards lower values) of divergence time (Fig. 2) or ancestral ϑ (Fig. S1) inferred from the sequences simulated by this program. Although the absolute values of the estimates may therefore be biased, the differences among models, number of loci, number of sequences and prior ranges should still be informative for the general trends discussed here.

SIMDIV was initially developed to produce data sets under the IM model and test the performance of a method estimating divergence parameters (Wang & Hey 2010). It has been further used in other studies, for example, to assess performances of different versions of IMa2 (Strasburg & Rieseberg 2010; Choi & Hey 2011), or to discriminate potential scenarios of divergence (Jackson & Austin 2012; Salas-Lizana *et al.* 2012). Given the differences highlighted in this study among simulation programs, we strongly recommend the comparison of multiple simulation programs in any study relying on simulated sequence data.

ESTIMATING PERFORMANCES OF THE IMA2 PROGRAM

Effect of model parameters

It appears that the best model conditions for estimating demographic parameters vary with the type of parameter (divergence time, migration rate or ϑ). Focusing on divergence time, the best estimates are associated with sequences simulated under a model that implements recent population divergence and no migration between populations (Model 3). In the absence of migration, poorer estimates seem to be associated with large divergence times (Models 1 and 2). However, while estimates under Model 2 have the largest CIs (almost similar to priors), they are occasionally very good. Because a large divergence time (relative to ϑ) is expected to result in complete lineage sorting between populations, it is possible that the level of shared ancestral polymorphism between the diverging populations is in fact informative for the purpose of inferring their time of divergence; thus, when complete lineage sorting is reached, which occurs with the highest probability in Model 2, the method is able to rely solely on the genetic distance separating the two populations to estimate the parameter, which could explain why the estimation is less accurate. Indeed, one of the benefits of the IM model is that it considers both gene flow and ancestral polymorphism to account for shared alleles among populations.

When migration does occur, estimating divergence time appears to be more difficult. With a recent divergence time (Models 3 and 4), the estimates appear in most cases widely different from the true value (2–3 times larger than the true value; Table 2). However, these estimates are associated with credible intervals spanning almost the entire prior range, so that the user can recognize that these estimates are unreliable. When the time of divergence is relatively large (Models 1 and 2), estimates are usually associated with shorter CIs, giving the illusion of better estimates, but are often far away from the true value. This is especially true for Model 2, for which the true value is systematically strongly underestimated, while in about

40% of the cases, even the range of the CI is far from the true value (upper limit set to ≤10 while the true value is 20).

In the absence of migration, estimation of migration rate (true value $m = 0$) seems slightly better when ϑ is large (Models 1 and 3) compared to the divergence time. While the migration parameter is clearly more difficult to estimate in the presence of migration, two optimistic observations can be drawn from our results. First, the true value is almost always included in the inferred credible interval, whatever the model considered, even in the presence of migration. Second, if the migration rate m is ≥0·05 (corresponding to a migration rate m of 0·00005 if the mutation rate μ is 10^{-3}), the method is capable in most cases of discriminating between a model with or without migration, since the credible interval usually does not include the value 0 where migration was implemented in the model.

In contrast with the divergence time parameter discussed above, ϑ is best estimated when the divergence time is long and ϑ is small (Model 2). In this case, the level of incomplete lineage sorting would indeed be expected to be irrelevant for the inference, since the estimated parameter characterizes the property of a single population, regardless of the state of the other. If the lineage sorting among populations is complete, which is more likely the case in Model 2 without migration than in other demographic conditions, the sequence diversity observed in one population is directly related to its parameter ϑ. If migration does occur, then the sequence diversity within one population also depends on what occurs in the other population. Because the divergence time is long, it should be easier to identify those sequences that are migrants than in the case of a recent split, and therefore to base the estimation only on those sequences that have not migrated from the other population. Note, however, that sequences generated under Model 2 do not allow for a better estimation of the migration rate (see above), which appears to contradict this explanation. In the case of Model 3, the occurrence of incomplete lineage sorting could be explained either by a recent divergence time (i.e. small relative to ϑ) or by migration. When migration occurs, we have seen that inferring migration rates becomes difficult. Thus, the method probably struggles with separating the two alternative interpretations. This may explain the less reliable estimates of ϑ when confronted with a relatively small divergence time. Finally, estimates of ϑ_{anc} are poor in general, probably revealing a general lack of information in sequence data for inferring this parameter, although larger divergence times relative to ϑ (Model 2 and 4) seem to induce better estimates for this parameter as well, probably for the same reason as those discussed for ϑ.

Overall, the performance of the IM model seems to depend on whether migration among the studied populations is strong. For example, someone studying two populations of a species characterized by a small ability to migrate and separated by a major geographic barrier (e.g. some species of snail located on two distant islands) could presume the absence of migration between them. In that case, relatively good estimates for the divergence time between the two populations can be expected, even more so with recent divergence times (i.e. relative to the population effective size). On the other hand, in the case of a much more mobile organism (e.g. some species of migratory birds), strong migration rates are expected, which will probably preclude the possibility to get a reliable estimate of divergence time, although population size, or even migration rates can probably still be somewhat estimated if larger data sets are used (see below).

Effect of sample size, number of loci and prior distributions

For any empirical phylogeographic study, because time and financial resources are limited, it is interesting to predict the additional information that can be gained by adding extra sequences or loci to an existing data set, or simply to determine the required number of sequences and loci for a new study, to answer a specific evolutionary question. We have found that the expected gain by adding sampled sequences or loci will vary greatly from case to case, and the best way to answer this question for a specific project is probably to perform a simulation study like ours with a demographic model designed specifically for that study. Nonetheless, some general trends can be derived from our results. Whatever the estimated parameter, if the credible interval around an estimate is large, increasing the number of sequences or loci should help to reduce its range. For divergence time and ϑ, it may not improve their accuracy however (because the same bias is maintained when increasing the number of loci and sequence). For migration rate, it should both reduce the range of the credible interval and improve the accuracy of the estimate (i.e. the CI will more likely include the true value). Overall, a data set of five loci appears sufficient for estimating demographic parameters in many situations, and increasing the number of loci beyond that point may not always be worth it. Note that Felsenstein (2006) has already investigated the impact of sample size and number of loci on parameter estimates, but under a simpler model of coalescence (single population). He inferred that the accuracy of maximum likelihood estimates increased rather slowly with sample size, while analysing multiple loci (rather than a single locus) strongly improved the estimates. In our study, increasing sample size from 10 to 40 sequences per population also improved the estimates strongly, probably due to the more complex model of population evolution that was investigated. Finally, we have found, not surprisingly, that increasing the range of the prior distribution of a parameter can increase the range of its posterior distribution, most conspicuously that of the estimated divergence time. In the cases we investigated here, it had no impact on the accuracy of the estimates however. It is then recommended to work with a small prior distribution, whenever possible (i.e. when historical information outside the analysed sequence data set is available).

Conclusion

Although it is clear in the light of our results that the best demographic conditions to estimate one parameter can differ widely from those to estimate another, our findings should be

useful for analysing empirical data. They first suggest the method is more appropriate to infer parameters from populations that remained completely isolated from each other since the split of their ancestral population, either because the organisms are characterized by a low capacity to disperse, or in the presence of a strong barrier to gene flow. Another important finding is that, in most demographic conditions (at least within the boundaries of the IM model), IMa2 appears reliable for discriminating between presence and absence of migration, although when migration does occur, estimating its extent should be more challenging. Other simulation studies have suggested that the same inference is more difficult when sequence data do not conform to the IM model, for example, when speciation takes place in an allopatric geographic setting with post-divergence gene flow or for a sympatric divergence with no current gene flow (Becquet & Przeworski 2009; Strasburg & Rieseberg 2011).

Finally, it is important to stress that all estimates presented in this study were inferred from an ideal situation, in which the model that generated the sequence data is identical to that generating the estimates. For this reason, the quality of these estimates should probably be considered as the best that can be obtained from DNA sequences, since empirical data are generated by much more complex systems. Becquet & Przeworski (2009) and Strasburg & Rieseberg (2010) have previously shown, even though they considered only a fraction of the parameter space investigated here, that deviating from the IM model when simulating the data can result in lower quality of estimates, depending on the model assumption being violated (with some deviations only slightly influencing the estimates). Assuming that the versions of the program IMa2 we used did not contain any major bug, we can consider the quality of the estimates presented here to reflect the limitation of the method in general and of the historical information contained in DNA sequence data.

Acknowledgements

We thank two anonymous reviewers and the associate editor for helpful comments on a previous version of our manuscript. MQ, FM and NE were supported by scholarships from the Belgian *Fonds pour la recherche scientifique* (F.R.S.-FNRS) and by the Van buuren Fund. PM is Research Associate at the F.R.S.-FNRS. Computational resources were provided by the High Performance Computing Centre, co-funded by both Free Universities of Brussels (ULB and VUB; HPC cluster 'Hydra').

Data accessibility

No empirical data were generated for this theoretical study, based exclusively on the analysis of computer-simulated sequences.

References

Abdo, Z., Crandall, K.A. & Joyce, P. (2004) Evaluating the performance of likelihood methods for detecting population structure and migration. *Molecular Ecology*, **13**, 837–851.

Becquet, C. & Przeworski, M. (2009) Learning about modes of speciation by computational approaches. *Evolution*, **63**, 2547–2562.

Bohonak, A.J. (1999) Dispersal, gene flow, and population structure. *The Quarterly Review of Biology*, **74**, 21–45.

Choi, S.C. & Hey, J. (2011) Joint inference of population assignment and demographic history. *Genetics*, **189**, 561–577.

Faubet, P., Waples, R.S. & Gaggiotti, O.E. (2007) Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Molecular Ecology*, **16**, 1149–1166.

Felsenstein, J. (2006) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691–700.

Hasegawa, M., Kishino, H. & Yano, T.A. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.

Hey, J. (2010) SIMDIV Manual [Documentation file]. Available with the program at http://genfaculty.rutgers.edu/hey/software

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. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 2785–2790.

Hudson, R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jackson, N. & Austin, C. (2012) Inferring the evolutionary history of divergence despite gene flow in a lizard species, *Scincella lateralis* (Scincidae), composed of cryptic lineages. *Biological Journal of the Linnean Society*, **107**, 192–209.

Kotlík, P., Deffontaine, V., Mascheretti, S., Zima, J., Michaux, J.R. & Searle, J.B. (2006) A northern glacial refugium for bank voles (Clethrionomys glareolus). *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 14860–14864.

Laval, G. & Excoffier, L. (2004) Simcoal 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. *Bioinformatics*, **20**, 2485–2487.

Neigel, J.E. (1997) A comparison of alternative strategies for estimating gene flow from genetic markers. *Annual Review of Ecology and Systematics*, **28**, 105–128.

Nielsen, R. & Wakeley, J. (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Nunes, M.D.S., Orozco-Ter Wengel, P., Kreissl, M. & Schlötterer, C. (2010) Multiple hybridization events between Drosophila simulans and Drosophila mauritiana are supported by mtDNA introgression. *Molecular Ecology*, **19**, 4695–4707.

Prada, C. & Hellberg, M.E. (2013) Long prereproductive selection and divergence by depth in a Caribbean candelabrum coral. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 3961–3966.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rambaut, A. & Grassly, N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences*, **13**, 235–238.

Salas-Lizana, R., Santini, N.S., Miranda-Pérez, A. & Piñero, D.I. (2012) The Pleistocene glacial cycles shaped the historical demography and phylogeography of a pine fungal endophyte. *Mycological Progress*, **11**, 569–581.

Slatkin, M. (1985) Gene flow in natural populations. *Annual Review of Ecology and Systematic*, **16**, 393–430.

Slatkin, M. (1987) Gene flow and the geographic structure of natural. *Science*, **236**, 787–792.

Strasburg, J.L. & Rieseberg, L.H. (2010) How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution*, **27**, 297–310.

Strasburg, J.L. & Rieseberg, L.H. (2011) Interpreting the estimated timing of migration events between hybridizing species. *Molecular Ecology*, **20**, 2353–2366.

Wang, Y. & Hey, J. (2010) Estimating divergence parameters with small samples from a large number of loci. *Genetics*, **184**, 363–379.

Whitlock, M.C. & McCauley, D.E. (1999) Indirect measures of gene flow and migration: $FST \neq 1/(4Nm + 1)$. *Heredity*, **82**, 117–125.

Received 22 August 2014; accepted 19 January 2015
Handling Editor: Emmanuel Paradis

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Fig. S1. Theta (ϑ) estimates obtained from DNA sequences simulated under Model 1, with and without migration, with three different simulation programs, for daughter (q) and ancestral (q_{anc}) populations.

Fig. S2. Migration rate estimates obtained from DNA sequences simulated under Model 1, with and without migration, with three different simulation programs.

Fig. S3. Estimates of different demographic parameters (A. standardized divergence time, B. standardized migration rate, C. ϑ, and D. ϑ_{anc}) obtained by analyzing DNA sequences (n = 40 per population) generated by simulations along the models depicted in Figure 1.

Fig. S4. Estimates of three demographic parameters obtained by analyzing DNA sequences from 5, 10 or 20 loci, generated by simulations along Model 3 without migration.

Fig. S5. Posterior distributions of estimated divergence time (t_0), provided for 5 different analyses (chosen randomly among 10–25 runs) for each of the three priors (p1, p1/2, p2.5) used when evaluating effect of the extent of the prior range on the estimation.

Fig. S6. Posterior distributions of estimated theta for the two daughter populations, q_0 and q_1, given for five different analyses (chosen randomly among the 10–25 runs) for each of the three priors (p1, p1/2, p2.5) used when evaluating effect of the extent of the prior range on the estimation.

Fig. S7. Posterior distributions of estimated theta for the ancestral populations, q_2, are given for five different analyses (chosen randomly among the 10–25 runs) for each of the three priors (p1, p1/2, p2.5) used when evaluating effect of the extent of the prior range on the estimation.

Fig. S8. Posterior distributions of estimated migration rate, here m_0, are given for five different analyses (chosen randomly among the 10–25 runs) for each of the three priors (p1, p1/2, p2.5) used when evaluating effect of the extent of the prior range on the estimation.

Table S1. Results from non-parametric multiple comparisons of IMa2 estimates obtained for datasets (n = 10 or 40) generated with the three different simulation programs (SIMDIV, Simcoal and ms) with the Kruskall–Wallis statistical test.

Table S2. Full results from IMa2 analyses on five loci datasets (n = 10 or 40 sequences) simulated under Model 1 with SIMDIV, Simcoal and ms a. without or b. with migration for sampling size of 10 (a. and b.) or 40 sequences (c.); all parameter estimates are given along with their estimated 95% highest posterior density interval, HPDHI (highest limit) and HPDL (lower limit).

Table S3. Full results from IMa2 analyses on five loci datasets simulated under Model 1–4 with SIMDIV a. without and b. With migration; all parameter estimates are given along with their estimated 95% highest posterior density interval, HPDHI (highest limit) and HPDL (lower limit).

Table S4. Full results from IMa2 analyses on various multilocus datasets (5, 10 or 20 loci) simulated with SIMDIV under Model 3 without migration; all parameter estimates are given along with their estimated 95% highest posterior density interval, HPDHI (highest limit) and HPDL (lower limit).

Table S5. Full results from IMa2 analyses on five loci datasets simulated with SIMDIV under Model 3, for which upper bound of the prior distributions is set to t_{max} = 100, ϑ_{max} = 100 and m_{max} = 1, or these maximum values divided by 2 or multiplied by 2.5; all parameter estimates are given along with their estimated 95% highest posterior density interval, HPDHI (highest limit) and HPDL (lower limit).

Table S6. Full results from IMa2 analyses on five loci datasets (n = 10, 40, or 100 sequences per population) simulated with Simdiv under Model 3; all parameter estimates are given along with their estimated 95% highest posterior density interval, HPDHI (highest limit) and HPDL (lower limit).

Table S7. Results from the Shapiro–Wilk normality test comparing estimates from datasets generated with different simulation programs (SIMDIV, Simcoal and ms) either on original estimates (c) or after square-root transformation (sq.r (c)).

Table S8. Results from the test of homogeneity of variances (Bartlett test) comparing estimates from datasets generated by different simulation programs (SIMDIV, Simcoal and ms), either conducted on original estimates or after a square-root transformation.