Molecular Ecology (2004) **13**, 827–836 doi: 10.1111/j.1365-294X.2004.02101.x

Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations

PETER BEERLI
Computer Science and Information Technology and Biological Sciences Department, Florida State University, Tallahassee FL 32306–4120 USA

Abstract

Current estimators of gene flow come in two methods; those that estimate parameters assuming that the populations investigated are a small random sample of a large number of populations and those that assume that all populations were sampled. Maximum likelihood or Bayesian approaches that estimate the migration rates and population sizes directly using coalescent theory can easily accommodate datasets that contain a population that has no data, a so-called 'ghost' population. This manipulation allows us to explore the effects of missing populations on the estimation of population sizes and migration rates between two specific populations. The biases of the inferred population parameters depend on the magnitude of the migration rate from the unknown populations. The effects on the population sizes are larger than the effects on the migration rates. The more immigrants from the unknown populations that are arriving in the sample populations the larger the estimated population sizes. Taking into account a ghost population improves or at least does not harm the estimation of population sizes. Estimates of the scaled migration rate *M* (migration rate per generation divided by the mutation rate per generation) are fairly robust as long as migration rates from the unknown populations are not huge. The inclusion of a ghost population does not improve the estimation of the migration rate *M*; when the migration rates are estimated as the number of immigrants Nm then a ghost population improves the estimates because of its effect on population size estimation. It seems that for 'real world' analyses one should carefully choose which populations to sample, but there is no need to sample every population in the neighbourhood of a population of interest.

Keywords: migration rate, gene flow, maximum likelihood, finite migration model, coalescence, population genetics, parallel execution, cluster

Received 18 July 2003; revision received 13 November 2003; accepted 13 November 2003

Introduction

When we study organisms in their natural habitat, we almost always need to discuss the relationship of the populations studied to each other or to populations that were ignored. The magnitude of exchange of genetic material between the populations whether, in the long run, they remain separate or fuse into a single population.

Correspondence: P. Beerli. CSIT, Dirac Science Library, Florida State University, Tallahassee FL 32306–4120 USA. Fax: USA-(850) 644 0098; E-mail: beerli@csit.fsu.edu

Researchers often use a measure of population divergence such as Sewall Wright's fixation index, F_{ST} (Wright 1937; Wright 1951), or similar statistics (Weir & Cockerham 1984; Michalakis & Excoffier 1996) that partition the genetic variance among individuals in a population and between populations. Most researchers want not only to know whether there is structure among populations but also about the magnitude of the gene flow among them. In conservation biology, the estimation of a migration rate between populations of an endangered species might even be the goal of the study. Wright (1951) showed that there is a direct relationship between F_{ST} and the magnitude of

gene flow between the populations for the *n*-island model. This approach, although widely used, is vulnerable to violations of its basic assumptions (Whitlock & McCauley 1999). It provides only an overall average of the migration rate assuming that the number of populations is very large, whereas a more detailed view is often needed. Researchers, ignoring the problem that there might be interdependence of more than just two populations, have used, and often misused, F_{ST}-based approaches to estimate pairwise migration rates from multiple population data. Fu *et al*. (2003) showed that for finite numbers of populations, this interdependence can be substantial. Several methods have recently been developed to take such interdependence into account. Nicholson *et al*. (2002) and Weir & Hill (2002) developed statistics to estimate an F_{ST}-analogue for each population, Wilson & Rannala (2003) and Pritchard *et al*. (2000) used allele frequencies to infer structure using Baysian approaches, whereas I and others (Beerli & Felsenstein 1999; Bahlo & Griffiths 2000; Nielsen & Wakeley 2001) target the direct estimation of underlying population parameters, such as the migration rate, using coalescence theory with maximum likelihood or Bayesian approaches. In all these new approaches it is assumed the sampled populations represent all the populations, in stark contrast to methods that estimate a single migration parameter assuming that the sampled populations are a random sample from a large number of populations (Rousset 1996; Wakeley & Aliacar 2001). Rarely, however, are all populations actually sampled. By neglecting populations that are not part of the sample, are the individual-population estimators giving a false impression of accuracy? Is it possible to estimate migration rates among arbitrarily chosen populations without the Herculean effort of sampling every population? This study explores the effect of missing data on gene flow analysis with my own maximum likelihood method MIGRATE (Beerli 2003), which is based on the coalescent. I expect that findings in this study are also valid for similar approaches, such as GENETREE (Bahlo, Griffiths 2000) and MDIV (Nielsen, Wakeley 2001).

Methods

The Markov chain Monte Carlo based coalescent methods (Wilson & Balding 1998; Beaumont 1999; Beerli & Felsenstein 1999; Bahlo, Griffiths 2000; Beerli & Felsenstein 2001; Nielsen, Wakeley 2001) allow for complicated population models, but it is commonly assumed that samples from all populations are in the dataset. This is rarely the case when researchers work with natural populations; some populations are not sampled because of logistic difficulties, or because one is not aware that additional populations exist. These statistical methods by themselves do not require that all sampled populations have data, but the effects of the unobserved populations on the results from these methods warrants exploration.

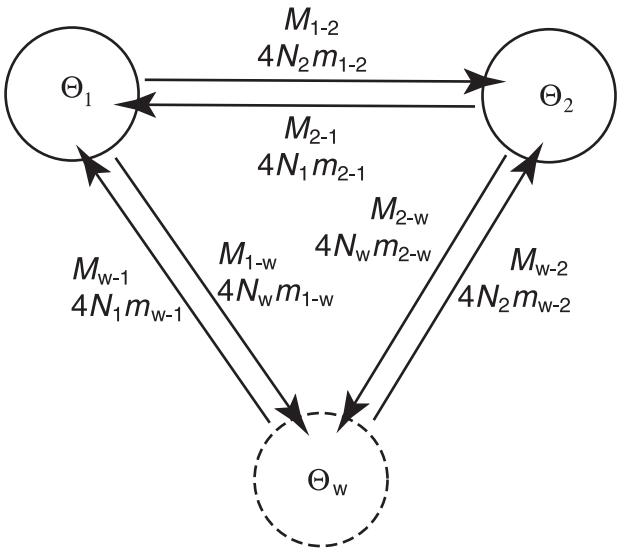

Fig. 1 Basic migration model used in the simulation study. Θ is $4\times$ effective population size $N_e\times$ mutation rate per generation and site; M is the scaled migration rate $m/$ where m is the immigration rate per generation. $4Nm$ (ΘM) is the number of immigrants per generation. The subscripts indicate the population and the direction of the migration, the first letter is the donor and the second the recipient.

Artificial datasets were created using a coalescent simulator similar in concept to Hudson's (1983) simulator (the C source code for the coalescent tree generator simtree and the data simulator simdata can be obtained by request from PB). The simulation study was kept as simple as possible to reduce problems caused by having too many parameters. I simulated two main scenarios to explore the effect of missing populations: the effect of the magnitude of migration, and the effect of the number of missing populations.

To explore the effect of the magnitude of the migration, a set of three interacting, equally sized populations was created of which only two are sampled (Fig. 1). I call the unsampled third population *world*. The immigration rate into the sampled populations from the *world* is called *world*-immigration to contrast it to the migration rates between the sampled populations (sample-immigration). Each dataset contains 20 individuals, 10 in each sampled population, each individual was scored for 100 unlinked loci and each locus has a length of 1000 bp. The populations all have the same size $\Theta = 0.01$, where Θ is $4\times$ effective population size $N_e\times$ mutation rate per generation and site; this is roughly equivalent to having 2500 individuals with a substitution rate of 10^{-6} per generation. For most of the simulations the migration rate between the sampled populations is kept at migration rate $M = 100$, where M

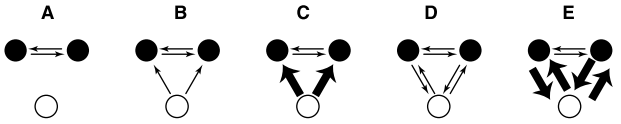

Fig. 2 Migration models used to generate the simulated datasets: unsampled *world* population are white, sampled populations, black. Thickness of arrows show the magnitude and direction of the immigration. Details are explained in the Methods section: (A) no immigration from *world*; (B, C) unidirectional immigration from *world*; (D, E) symmetric immigration.

is the immigration rate m scaled by . Using the above substitution rate, $M = 100$ translates into $m = 0.0001$ per generation. This is equivalent to one immigrant every four generations ($4N_e^{(1)}m = 4N_e^{(2)}m = 1$). The migration rate into population 1 from 2 is labelled $M_{2\text{-}1}$ and the migration from *world* to 1 similarly is $M_{\text{world-1}}$.

The following scenarios were simulated:

No immigration from the world: (Fig. 2A) the two sampled populations do not receive migrants from the *world* population.

Unequal world-immigration (Fig. 2B,C): (B) the *world*-immigration is of the same magnitude as the migration rate between the two populations, $M = 100$ ($4N_e^{(i)}m_{\text{world-1}} = (1)$; (C) the immigration from the *world* is much bigger than the exchange between the samples, $M = 1000$ ($4N_e^{(i)}m_{\text{world-1}} = 10$). There is no migration from the sample to the *world*.

Symmetric immigration from the world (Fig. 2D,E): All migration rates are the same as before except that all immigration from the *world* populations are matched with an immigration into the *world* population with the same magnitude.

To check the variability of the outcome, I generated 10 additional datasets for the scenario with symmetric migration rates between all three populations (Fig. 2D), and two further datasets containing 50 and 100 individuals using the same scenario.

For the analysis of the effect of the number of missing populations, datasets were created in which two sampled populations are part of a network of unobserved populations; sets of 3, 5, and 9 populations were simulated. For these simulations I kept the migration rate the same between all populations (Fig. 2D).

Additionally, effects of the *world* populations when there is no migration between the sample populations ($M = 0$) were analysed, using unequal and symmetric *world*-immigration scenarios.

I analysed these artificial datasets with MIGRATE (Beerli and Felsenstein 1999; Beerli, Felsenstein 2001). MIGRATE estimates migration rates and population sizes jointly using a maximum likelihood approach that is based on the coalescent (for an overview see Kingman 2000). The program finds parameters by maximizing a relative likelihood using the Markov chain Monte Carlo sampling scheme devised by Metropolis *et al*. (1953) with modifications by Hastings (1970). For a set of n populations, n population sizes and $n(n-1)$ immigration rates are estimated. The population sizes are reported as Θ (that is, $4N_e$ for nuclear data). The immigration rates are estimated as M, which is $m/$.

Each dataset was analysed by assuming first that there are only two populations and second, that there is in addition a third unknown population, which I call a ghost population because the data do not indicate its presence or absence or how many such populations are present. For datasets with multiple unknown populations, only one ghost population was used. For each dataset, I performed two MIGRATE runs to get a rough estimate of the Markov chain Monte Carlo error. The design of the study allowed to employ an ANOVA analysis that checked the major effects of the variance introduced by having two datasets per scenario, the two replicates, the magnitude of the *world*-immigration, and the variance of the sample-immigration rates. The ANOVA analysis was performed with MATHEMATICA 4.2 (Wolfram Research 1999). Mean square errors (MSE) were calculated to explore the benefits of including the ghost population in the analysis. The MSE measure is the expectation of the squared differences of the estimates and the expected values, $\text{MSE} = E(W - T)^2 = Var(W) + Bias^2$, where W is the estimate and T is its expectation. Here, T is the values used to simulate the dataset (the truth), and the *Bias* is the difference from the mean estimate and the truth. The MSE incorporates two components, variance and bias. One seeks the method that minimizes MSE.

The program version used was MIGRATE 1.7.3 (Beerli 2003) and all options were default except that a heating scheme (Markov Coupled Markov chain Monte Carlo: Geyer & Thompson 1992) with three heated chains and one cold chain in effect. The program was compiled for parallel execution using the Message Passing Interface (Gropp *et al*. 1999) and the analyses were done on an IBM eServer pSeries 690 cluster running AIX and using up to 101 processors concurrently. Details of the parallel implementation and its performance are described in the Appendix.

Results

Effects on the estimation of the population size

In Fig. 3 results from the three scenarios each with two datasets generated with the same parameters are shown. The migration rate is unidirectional from the *world* to the sampled populations (Fig. 3 A1 and A2). When there is no connection between the *world* and the samples, the two-population analysis estimates the sizes for the two populations well, with ($\text{MSE} = 0.088 \times 10^{-6}$) with an expected

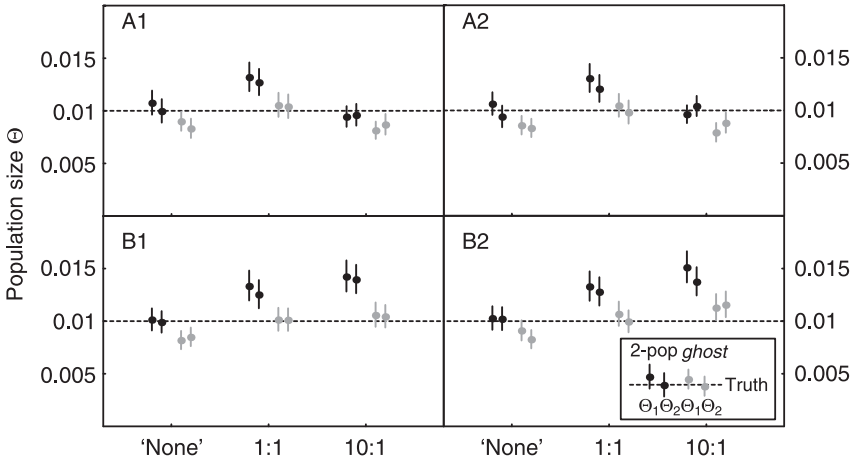

Fig. 3 Population size estimates with and without *ghost*-analysis. A1 and A2 are results from two simulated datasets using unequal migration rates between *world* and samples. B1 and B2 are results from datasets using symmetric migration rates between *world* and samples. Dots mark the maximum likelihood estimate, the lines cover the range of the approximate 98% support interval. The box in the corner indicates the order of the parameters in the graph: each subblock consists of the sizes Θ_1 and Θ_2 using only the sampled populations (black), including a *ghost* (grey), respectively. The groups 'none', 1:1; 10:1 indicate the strength of the immigration from the unsampled *world* population, where for example 10:1 means that the immigration from the *world* was 10 times the immigration rate between the sampled populations.

Table 1 Mean square error (MSE) of the different scenarios (*world* population is source: B, C, and symmetric rates: D, E) used to analyse the simulated datasets. The MSE and its standard deviation is calculated based on the average of parameters over two replicates (numbers in parantheses show the values for 10 replicates). The scenarios A–E are the same as in Fig. 2. 2:3 and 2:7 are the ratios between sampled and unsampled populations. The true values for Θ and *M* are 0.01 and 100, respectively

	Mean square error Θ		Mean square error *M*	
Simulated scenario	Two-population analysis (×10⁶)	*ghost*-analysis (×10⁶)	Two-population-analysis	*ghost*-analysis
A No *world*-immigration	0.088 ± 0.022	1.73 ± 0.26	115 ± 105	286 ± 325
B Medium *world*-immigration	8.2 ± 0.03	0.15 ± 0.05	190 ± 2	4 ± 5
C High *world*-immigration	0.33 ± 0.09	2.67 ± 0.05	45 041 ± 12 562	30 637 ± 7226
D Medium *world*-immigration	9.02 ± 0.94	0.01 ± 0.00	320 ± 101	517 ± 110
	(9.46 ± 2.89)	(0.10 ± 0.16)	(125 ± 144)	(144 ± 157)
E High *world*-immigration	15.6 ± 1.25	0.86 ± 0.89	20 862 ± 5791	15 068 ± 7651
2:3	65.10 ± 79.04	15.19 ± 2.13	58 ± 19	232 ± 202
2:7	647.01 ± 574.98	186.54 ± 23.50	40 ± 37	4 ± 2

value of Θ = 0.01. Both datasets recover the true value whereas the *ghost*-analysis tends to underestimate the population size (MSE = 1.73×10^{-6}). With increased *world*-immigration, the two-population analysis overestimates the sizes considerably (MSE = 8.2×10^{-6}), whereas the *ghost*-analysis recovers the true values (MSE = 0.15×10^{-6}). The result for the high *world*-immigration is puzzling as one might expect that all the parameters might be overestimated, but the two-population analysis recovers the true values with an MSE of 0.33×10^{-6}. A summary of the MSEs for all parameters is shown in Table 1. The results for Θ from the simulations with symmetric immigration rates between sampled populations and the *world* reveal s a similar pattern except for the high migration scenario (Fig. 3B1 and B2). With low *world*-immigration, the two-population analysis overestimates and the *ghost*-analysis recovers the true values, but the high *world*-immigration simulations show over-estimation of the population sizes. There is a tight correlation with the magnitude of *world*-immigration rates: the MSE for all three two-population scenarios are 0.09×10^{-6}, 9.0×10^{-6}, 15.6×10^{-6}. The bias of the *ghost*-analysis is less well correlated with the magnitude of the immigration with MSEs of 1.7×10^{-6}, 0.01×10^{-6}, and 0.86×10^{-6}.

Effects on the estimation of the sample-immigration rates

The effect on the immigration rate *M* between the samples follows a pattern very similar to those described for the

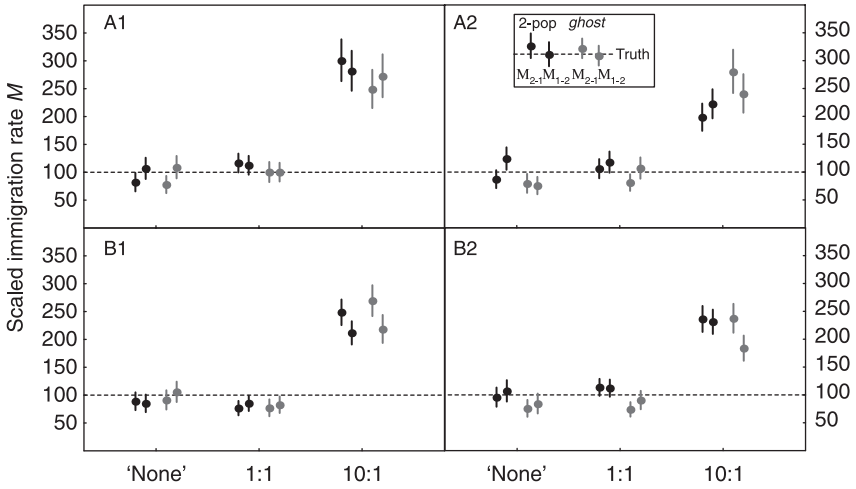

Fig. 4 Immigration rates estimates with and without *ghost*-analysis. A1 and A2 are results from two simulated datasets using unequal migration rates between *world* and samples. B1 and B2 are results from datasets using equal migration rates between word and samples. Dots mark the maximum likelihood estimate of M, which is migration rate m over mutation rate , the lines cover the range of the approximate 98% support interval. The box in the corner indicates the order of the parameters in the graph: each sub-block consists of the sizes $M_{2\text{-}1}$ and $M_{1\text{-}2}$ using only the sampled populations (black), including a *ghost* population (grey), respectively. The groups 'none', 1:1; 10:1 indicate the strength of the immigration from the unsampled *world* population, where for example 10:1 means that the immigration from the *world* was 10 times the immigration rate between the sampled populations.

effective population size Θ. The two sample-immigration rates ($M_{1\text{-}2}$, $M_{2\text{-}1}$) should be equal, even when the *world*-immigration is unequal, because the immigration pattern between the samples was kept symmetrical. The $M_{1\text{-}2}$ and $M_{2\text{-}1}$ are quite accurate when there is no or low immigration from the *world* and overestimated with *world*-immigration (Fig. 4). The MSE for the two-population and the *ghost*-analysis are very similar: for some scenarios the two-population analyses performs better than the *ghost*-analyses, but there is no clear pattern emerging except that MSE for the two-population analysis is correlated with the magnitude of *world*-immigration (Table 1).

An ANOVA analysis of the major effects of the data used for Fig. 4(B) was performed. We compared the effects of the analysis-type (Two-population vs. *ghost*-analysis), the magnitude of *world*-immigration (none, 1:1, 10:1; Fig. 2 A,D,E), the variance between datasets, the variance between replicates of the same dataset (the variance resulting from the Markov chain Monte Carlo procedure), and the variance of the sample-immigration rates within a replicate. At the significance level $\alpha = 0.05$ and using a Bonferroni test only two effects are significant: the analysis-type (there is a difference whether one is using a *ghost* population or not), and the magnitude of *world*-immigration (the results for the high *world*-immigration are very different from the others). Upon closer inspection, the effect of the analysis-types can be differentiated because the group 1:1 shows that the *ghost*-analyses are much more biased than the two-population analysis. This case was re-examined with 10 simulated datasets using the same parameter settings as

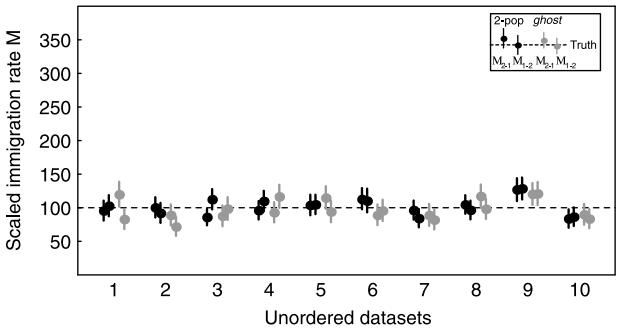

Fig. 5 Immigration rate estimates of 10 independent datasets for 100 loci, simulated using a value of $M = 100$ for all migration rates. Dots mark the maximum likelihood estimate, the lines cover the range of the approximate 98% support interval. The box in the corner indicates the order of the parameters in the graph: each sub-block consists of the sizes $M_{2\text{-}1}$ and $M_{1\text{-}2}$ using only the sampled populations (black), including a *ghost* population (grey), respectively.

the two datasets used for the ANOVA. Visual inspection of the results does not reveal a consistent difference between the two-population and the *ghost*-analysis (Fig. 5). Averaging over the 10 different datasets gives parameter values that are very close to the truth (Table 2). The MSE of the immigration rate based on these 10 datasets for the two-population analysis is 125, and 144 for the *ghost*-analysis, somewhat better numbers than the crude MSE from two data sets of 320 and 517. The smaller the MSE, the more confidence we have that the method has a small bias and small variance. The variances for the sample immigration

Table 2 Averages of the migration rate M over 10 datasets. M is $m/$ where m is the migration rate per generation and is the mutation rate per generation and site. The datasets were simulated with migration rates from population two into population one ($M_{2\text{-}1}$) and from one into two ($M_{1\text{-}2}$) of 100, the migration rate from and to the *world* population was also set to 100. The percentage values are the averages of the respective percentiles, MLE is the average of the maximum likelihood estimates

Immigration rate		Two-population-analysis	*ghost*-analysis
$M_{2\text{-}1}$	1%	86	85
	MLE	100	100
	99%	116	117
$M_{1\text{-}2}$	1%	88	79
	MLE	102	94
	99%	118	110

rates of the 10 replicates are 137.4 and 150.8, respectively. The *ghost*-analysis has higher variance because it estimates nine parameters instead of just four; for low *world*-immigration rates both estimators seem to have small bias because the variance is of the same magnitude as the MSE.

No direct migration between the sampled populations

Table 3 shows the MSEs for datasets where there is no migration between the sample populations. All gene flow between the samples is indirect through the *world* population. The two-population analyses have larger MSEs for Θ and M than the *ghost*-analyses, except for the unidirectional, large *world*-immigration where the two-population has a minimal MSE for Θ.

Effects of the number of missed populations

The estimated sample-immigration rates M are astonishingly robust when we increase the number of unsampled populations (Fig. 6, bottom). The estimates for ratios of 2:1 between sampled and unsampled populations, 2:3, and 2:7 are very similar for both types of analyses, both the two-population and the *ghost*-analysis seems to improve with more unsampled populations (Table 1). In contrast, the

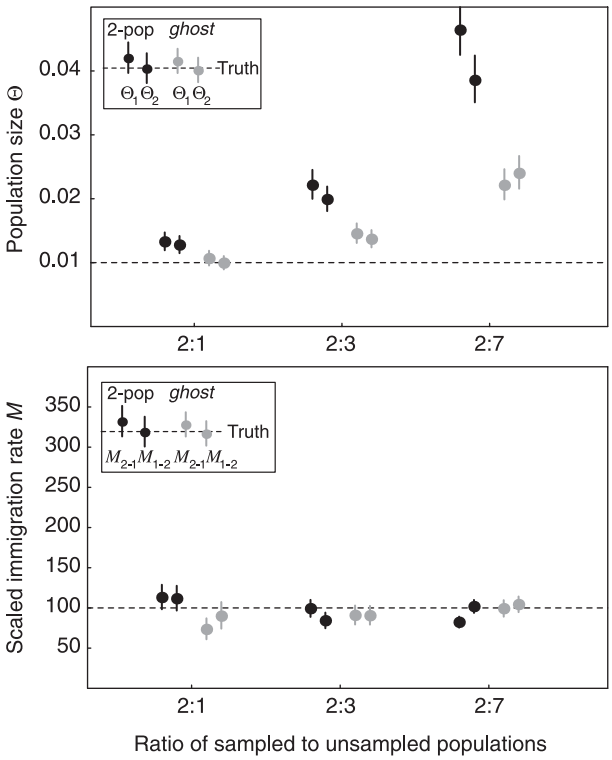

Fig. 6 Effect of the number of unsampled vs. sampled populations. Dots mark the maximum likelihood estimate for population sizes (top) and the immigration rates (bottom), the lines cover the range of the approximate 98% support interval. The box in the corner indicates the order of the parameters in the graph: each sub-block consists of the sizes $M_{2\text{-}1}$ and $M_{1\text{-}2}$ using only the sampled populations (black), including a *ghost* population (grey), respectively. The groups 2:1, 2:3, 2:7 reflect the ratio of sampled to unsampled populations.

Table 3 Mean square error (MSE) when the migration between the samples is zero and migrants are only exchanged through the *world* population. The MSE and its standard deviation are calculated based on the average of parameters over two replicates. Letters B-E mark the scenarios B–E from Fig. 2, except that the true values for M between the samples are 0

	Mean square error Θ		Mean square error M	
Simulated scenario	Two-population analysis (×10⁻⁶)	*ghost*-analysis (×10⁻⁶)	Two-population analysis	*ghost*-analysis
B Medium *world*-immigration	8.42 ± 0.78	0.37 ± 0.02	1547 ± 72	658 ± 112
C High *world*-immigration	0.18 ± 0.02	2.87 ± 0.58	83 918 ± 1982	65 655 ± 234
D Medium *world*-immigration	9.60 ± 2.220	0.20 ± 0.12	1154 ± 142	828 ± 106
E High *world*-immigration	21.39 ± 1.77	1.87 ± 0.52	38 381 ± 5404	39 628 ± 2881

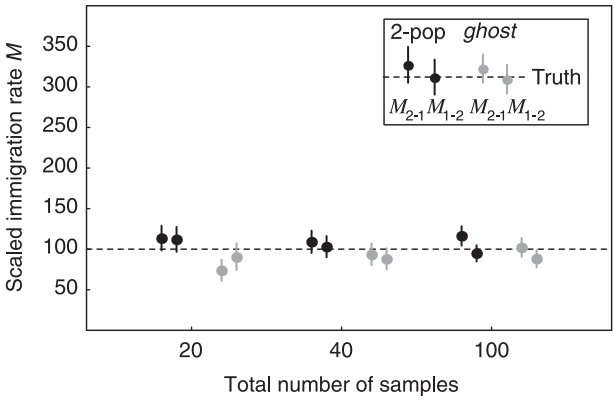

Fig. 7 Comparison of the effect of sample size on the accuracy of the migration rate estimate. M is $m/$, where m is the immigration rate per generation and is the mutation rate per generation. Dots mark the maximum likelihood estimate, the lines cover the range of the approximate 98% support interval. The box in the corner indicates the order of the parameters in the graph: each sub-block consists of the sizes $M_{2\text{-}1}$ and $M_{1\text{-}2}$ ignoring the *world* (black), taking the *world* into account (grey), respectively.

estimates of the population sizes are strongly affected by missed populations (Fig. 6, top). For the two-population analysis, the MSE increase about nine fold between the ratios of sampled vs. missing populations of 2:1, 2:3, and 2:7. For the *ghost*-analysis, the values are smaller, but they show a similar, but weaker, trend as the two-population analysis (Table 1).

Sample size

To make sure that the above comparisons are not an effect of small sample size (10 individuals per population), additional datasets with 50 and 100 total individuals were analysed. These datasets were simulated under the scenario that all migration rates are equal (Fig. 2D). The effect of small sample size on the estimates of the immigration rate is minimal (Fig. 7), except that the support intervals are somewhat larger with fewer individuals. The results for the population sizes are not shown but follow the same pattern; this is expected and has been found before (Pluzhnikov & Donnelly 1996).

Effects on the estimation of the number of immigrants into a sample population

The number of immigrants can be expressed as $4N_em = \Theta M = 4N_e \times m/$. The biases that affect Θ and M will affect the $4N_em$ the same way. The biases of Θ for scenarios with medium and high world-immigration rates are large whereas the biases for M are much smaller for medium *world*-immigration than for large world-immigration, the biases for $4N_em$ follow closely the biases for Θ.

Discussion

Results were partly expected and partly surprising. The estimates of the migration rate M between the samples are stable when there is no or low immigration from the *world*. The estimates deteriorate with large immigration rates from the *world*: the many alleles imported into the two sampled populations from the same source increase the estimates of M because the occurrence of the same allele in all the populations increases the chance that there was a migration. Surprisingly, with moderate *world*-immigration rates the number of missed populations does not affect the estimates of the migration rates between the samples. The local alleles are not swamped by alleles from *world* as in the large *world*-immigration scenarios and therefore allow to estimate successfully the migration rates between the samples. This confirms Hudson's (1998) assertion that is possible to estimate a migration rate in a subgroup of populations in a n-island model based on a 'local' F_{ST}. Whether one uses a *ghost* population in the analysis or not, is not that important because ignoring unknown populations or taking them into account provides very similar results for the migration rate estimates.

The estimates of population sizes show more deviation from the true values than those for the migration rates. The quality of the estimates depends on the simulation scenario. With symmetric migration rates between the sample and the *world* a steady increase in population sizes was found and the deviation from the true value gets larger when the immigration rate increases or the number of missed population increases; having a migration rate M of 1000 from a single missed population is similar of having an M of 100 from 10 missed populations. The *ghost* population analysis improves the estimates of Θ somewhat with small immigration rates but the estimates are still biased upward when the gene flow from the unknown populations is large. There is a discrepancy between the source-sink and the symmetric migration scenario. In the source-sink scenario with large immigration rate the population sizes are estimated quite accurately. This is an artifact because the sample populations get swamped with alleles from the *world* and essentially become copies of the *world* population and therefore should have its size; instead of measuring the size of the samples under these conditions one is measuring the size of the unknown population. With symmetric and high migration rates all populations exchange many migrants through the *world* populations, so that the three populations essentially behave like one large population in which the coalescences are independent from the location of the sampling (Nagylaki 2000).

Fu *et al*. (2003) showed that many estimators using allele frequencies have difficulties estimating the degree of isolation for a finite number of small connected populations because the allele frequencies co-vary. This effect is more

pronounced with small sizes because genetic drift acts on all populations and on the group as a whole. One might wonder how much approaches based on the limit of infinite number of populations (for example Wright 1951; Wakeley, Aliacar 2001) correctly estimate migration rates when the immigration rates deviate from an *n*-island model. I expect that for moderate and low migration rates the overall parameter estimates might be quite accurate because there was no strong deviation in simulations for these scenarios even with a large number of unsampled populations, but one would need to evaluate the behaviour of these methods when the immigration rates from *world* populations to the sample populations are large.

Migration rates might be expressed as genetic distances that take into account variances and covariances of allele frequencies (Wood 1986; but see Fu *et al*. 2003). When there is no direct migration between the sample populations, variability still can be distributed through the *world* and we would expect an upwards bias for the sample migration rates because with large *world*-immigration the sample populations are swamped with alleles from the *world*. This results in a considerably biased sample-immigration rate for the scenario with unidirectional *world*-immigration where there is no exchange between the sample populations, the many similar *world*-alleles make it impossible to establish whether this is the result of sample-immigration or export of *world*-alleles into the samples. When there is more interest in historical processes than interest in variability patterns this distinction is relevant and migration rates cannot be replaced by pairwise genetic distances. Inclusion of a *ghost* population seems to improve the estimates of sample population sizes considerably, suggesting that a simple pairwise treatment of migration incorporates some of the variability imported from locations other than the pair under consideration and so will lead to overestimation of local variability.

It seems unnecessary to add a *ghost* population to analyse migration rates *M* because in many comparisons the MSE do not strongly favour the *ghost*-analysis over a two-population approach although it seems that the two-population method is favoured because of the larger variance of the *ghost*-analysis caused by the higher number of parameters to estimate. If the migration rate *M* from the unsampled population is known to be large or if the focus of the analysis is to estimate the population size Θ, or the number of immigrants $4N_em$, which is the product of Θ and *M*, then the addition of a *ghost* does help to reduce the upwards bias. Bittner & King (2003) were using my *ghost* approach to estimate $4N_em$ between snake populations on islands in Lake Erie. They report that a *ghost* is useful only when few populations were sampled but when additional samples from more populations are available the inclusion of a *ghost* has no benefit for the estimation of $4N_em$. We might extrapolate that when only two populations are sampled, the population sizes are most likely overestimated and the only hope for getting accurate numbers is to sample the dominating populations. Adding samples from other populations is fairly simple because one is not required to sample huge numbers of individuals to get decent results from a coalescent based analysis.

Acknowledgements

This study was supported by a grant from the National Science Foundation DEB-0108249 to Scott Edwards and the author; preliminary simulations were initiated while I was at the University of Washington supported by grants from National Science Foundation (DEB-9815650), the National Institute of Health (GM-51929 and HG-01989) to Joseph Felsenstein, whom I thank for many discussions on the topic. The simulations were supported by Florida State University School for Computational Science and Information Technology and utilized their IBM eServer pSeries 690 Power4-based supercomputer *Eclipse*. I also want to thank Thomas Uzzell, Laurent Excoffier, Mary Kuhner, Scott Edwards, Richard King, and an anonymous reviewer for helpful comments on the manuscript.

References

Bahlo M, Griffiths RC (2000) Inference from gene trees in a subdivided population. *Theoretical Popul Biology*, **57**, 79–95.

Beaumont MA (1999) Detecting population expansion and decline using microsatellites. *Genetics*, **153**, 2013–2029.

Beerli P (2003) MIGRATE — a maximum likelihood program to estimate gene flow using the coalescent, Tallahassee/Seattle. http://evolution.gs.washington.edu/lamarc/migrate/html

Beerli P, Felsenstein J (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences of the USA*, **98**, 4563–4568.

Bittner TD, King RB (2003) Gene flow and melanism in garter snakes revisited: a comparison of molecular makers and island vs. coalescent models. *Biology Journal of the Linnean Society*, **79**, 389–399.

Fu R, Gelfand AE, Holsinger KE (2003) Exact moment calculations for genetic models with migration, mutation, and drift. *Theoretical Population Biology*, **63**, 231–243.

Geyer CJ, Thompson EA (1992) Constrained Monte-Carlo maximum-likelihood for dependent data. *Journal of the Royal Statistical Society Series B-Methodology*, **54**, 657–699.

Gropp W, Lusk E, Skjellum A (1999) *Using MPI Portable Parallel Programming with the Message–Passing Interface*, 2nd edn. MIT Press, Cambridge, Mass.

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183–201.

Hudson RR (1998) Island models and the coalescent process. *Molecular Ecology*, **7**, 413–418.

Kingman JF (2000) Origins of the coalescent 1974–82. *Genetics*, **156**, 1461–1463.

Metropolis N, Rosenbluth AW, Rosenbluth N, Teller AH, Teller E (1953) Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics*, **142**, 1061–1064.

Nagylaki T (2000) Geographical invariance and the strong-migration limit in subdivided populations. *Journal of Mathematical Biology*, **41**, 123–142.

Nicholson G, Smith AV, Jonsson F *et al*. (2002) Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, **64**, 695–715.

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

Pluzhnikov A, Donnelly P (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics*, **144**, 1247–1262.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation processes. *Genetics*, **142**, 1357–1362.

Wakeley J, Aliacar N (2001) Gene genealogies in a metapopulation. *Genetics*, **159**, 893–905.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Weir BS, Hill WG (2002) Estimating F-statistics. *Annual Review of Genetics*, **36**, 721–750.

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: F_{ST} not equal to $1/(4Nm + 1)$. *Heredity*, **82** (2), 117–125.

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. *Genetics*, **150**, 499–510.

Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, **163**, 1177–1191.

Wolfram Research (1999) Mathematica: Wolfram Research, Inc, Champaign, Illinois.

Wood JW (1986) Convergence of genetic distances in a migration matrix model. *American Journal of Physical Anthropology*, **71**, 209–219.

Wright S (1937) The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the USA*, **23**, 307–320.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Appendix

The program MIGRATE-N summarizes multiple unlinked loci calculating the likelihood

$$L(\Theta, M) = \sum_{l=1}^{loci} \int_{G} \Pr(G \mid \Theta, M) \Pr(D_l \mid G) dG$$

(Beerli & Felsenstein 1999). Each locus is independent from any other so that the integration over all possible genealogies for each locus can be run independently. This makes the problem embarrassingly parallel. On multiple computers one can run all loci concurrently, and reduce the analysis time considerably. MIGRATE-N can be compiled for parallel machines utilizing MPI (Gropp *et al*. 1999). The current version uses a master-worker architecture. The flow of the analysis is as follows: the parameter file is read by the master-node. On interactive systems the menu can be displayed (all input/output related function are guided through the master). After the menu, the data are read and distributed to all worker nodes; the master orchestrates the workers, each of which gets a locus to work on. Once a locus is finished, the worker receives either a new locus or waits until all other workers are done with their work; the master then calculates the maximum likelihood estimate (MLE) by delegating the calculation of likelihoods and gradients to the workers. When the MLE is found, the first overview table is printed and the workers send all their locus summary data (sampled genealogies) to the master so that they can be redistributed to all other workers. After the redistribution of the data, the workers calculate the approximate support intervals for each parameter using the method of profile likelihood. The results are then forwarded to the master and printed in the outfile. The program needs to run on minimally two nodes and maximally as many as one can accommodate. A natural upper limit is the maximum number of loci or of parameters.

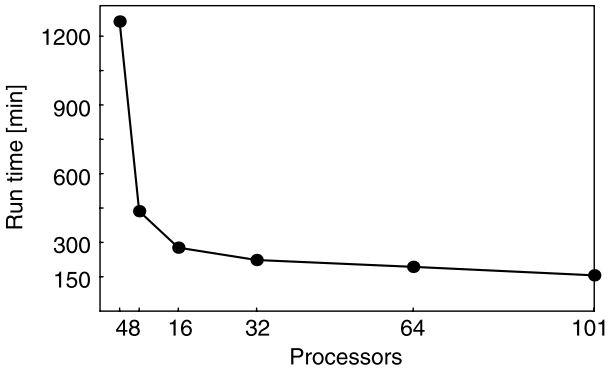

Fig. 8 Comparison of the run-time improvement of the parallel version of MIGRATE-N.

A typical speedup is displayed in Fig. 8 and it shows that for a dataset with 100 loci and nine parameters, 32 processors are very efficient and the use of more processors does not greatly improve the speed, although the 101 processor run is still 1.4 times faster than the 32-processor run. Even so the program is 'embarrassingly parallel'; with more nodes more data need to be transferred on the network, which is much slower than the central processing unit (CPU). Another problem is that work on some loci is much faster than work on others; if all the k loci are distributed to k nodes then for further computation one needs to wait for that node that received the locus that was most time consuming to compute. When each node can take several loci then some loci will be calculated rapidly and others slowly, averaging the total waiting time.

With the help of a computer-savvy person it is feasible for a lab group to set up a small cluster or group of connected workstations, and run batch jobs of MIGRATE-N without blocking individual researcher's desktop computers and get a decent turn-around time for individual runs of the program.