Evaluating the performance of likelihood methods for detecting population structure and migration

ZAID ABDO,*† KEITH A. CRANDALL‡ and PAUL JOYCE*†
*Department of Mathematics, PO Box 441103, University of Idaho, Moscow, ID 83844–1104 USA, †Initiative in Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, USA, ‡Department of Integrative Biology, Brigham Young University, Provo, UT 84602, USA

Abstract
A plethora of statistical models have recently been developed to estimate components of population genetic history. Very few of these methods, however, have been adequately evaluated for their performance in accurately estimating population genetic parameters of interest. In this paper, we continue a research program of evaluating population genetic methods through computer simulation. Specifically, we examine the software MIGRATE-N 1.6.8 and test the accuracy of this software to estimate genetic diversity (Θ), migration rates, and confidence intervals. We simulated nucleotide sequence data under a neutral coalescent model with lengths of 500 bp and 1000 bp, and with three different per site Θ values of (0.00025, 0.0025, 0.025) crossed with four different migration rates (0.0000025, 0.025, 0.25, 2.5) to construct 1000 evolutionary trees per-combination per-sequence-length. We found that while MIGRATE-N 1.6.8 performs reasonably well in estimating genetic diversity (Θ), it does poorly at estimating migration rates and the confidence intervals associated with them. We recommend researchers use this software with caution under conditions similar to those used in this evaluation.

Keywords: coalescent, likelihood, migrate, migration, population structure

Introduction
Effectively determining population substructure and levels of gene flow is an important aspect of population genetics. Such information is essential for accurate estimates of effective population sizes, genetic diversity, and migration rates — all key parameters in conservation biology, molecular ecology (especially metapopulation analyses), and population genetics. The standard approach to measuring population structure is through the use of \( F \)-statistics (Wright 1951, 1965; Neigel 2002). However, recent population genetic theory, namely the coalescent (Tavaré 1984; Hudson 1990), has allowed for the development of more sophisticated measures of population structure that take into account more of the underlying biology of populations and thereby produce more information on population structure than values from \( F \)-statistics (Pearse & Crandall 2004). The coalescent process provides a framework for likelihood based statistical analysis with great potential for analysing DNA sequence data that arise in population genetics. Although the number of these methods increased during the last few years, not much work has been done to test their accuracy and their relative performance. In this paper we continue an effort (Posada & Crandall 2001), and (Brown et al. 2001) aimed at evaluating methods for detecting and estimating historical demographic events and population genetic parameters.

There are two standard approaches to assess method performance and compare performance with other methods. The first is by analysing empirical data were the true underlying history of the population is assumed to be known and results are compared to this assumption. This approach obviously has the drawback that that assumption of underlying truth may be incorrect. It does, however, have the advantage of using real data. The second approach is to simulate data, thereby allowing the researcher to know the truth and compare with results from different approaches. The drawback here is that simulated data tend to be much more simplistic (and therefore easier to analyse)
than real data (Brauer et al. 2002). We take the second approach to validate the effectiveness of programs aimed at detecting population structure. Specifically in this paper, we evaluate the performance of the software migrate-n version 1.6.8 (Beerli & Felsenstein 1999, 2001; Beerli 2002) in its ability to accurately estimate genetic diversity and migration rates. We refer to this version of migrate-n as migrate throughout.

Theory and background

migrate attempts to estimate the migration rate between multiple populations using a maximum-likelihood and coalescent-theory approach (Beerli & Felsenstein 1999, 2001; Beerli 2002). In the next two subsections we will outline our general coalescent model and then describe the model used by migrate in particular. We will then outline the general approach taken in our simulation studies.

The coalescent

The neutral coalescent has been the central focus of much of theoretical population genetics for the last 20 years (Tavaré 1984; Hudson 1990; Nordborg 2001) due to the practical insights and relative mathematical elegance that comes from examining gene genealogies as they go back in time (coalesce). In the neutral coalescent, individuals are equally likely to reproduce within a population. This enables the separation between the neutral mutation process and the genealogical process. Hence, the history of evolution can be built by first constructing the genealogy and then simulating the mutations using an appropriate model of evolution (Neuhauser 2001; Nordborg 2001).

The two components of the genealogy are the topology and the branch lengths. The topology is built by modelling the genealogy of a sample of individuals backward in time regardless of the rest of the population. Going backward in time, and assuming selective neutrality and a Wright-Fisher model, individuals pick their parents independently at random from the previous generation with probability 1/(2N), where N is the effective population size. Individuals coalesce whenever they pick the same parent.

The length of a branch represents the waiting time until a coalescence event occurs. The coalescent approximates the distribution of this waiting time, scaled by twice the effective population size 2N, to an exponential distribution with mean 2/(k(k − 1)), where k is the number of currently available lineages. This approximation holds as long as 2N is reasonably large (Neuhauser 2001; Nordborg 2001).

In the case of the coalescent-with-migration (structured coalescent), individuals need to be in the same subpopulation (or patch) to coalesce. Hence, the probability of two individuals coalescing will depend on the rate of migration between the different subpopulations. Two types of events occur under this model: migration and coalescence. It can be shown that in the limit (as the total effective population size goes to infinity) and under some regularity conditions (Neuhauser 2001; Nordborg 2001; Felsenstein 2003) that the waiting time until the first event (migration or coalescence) is exponentially distributed with a rate equal to the sum of the rates of all possible events. Following Nordborg (2001), this rate is represented in the following equation:

\[ h(k_1, k_2, \ldots) = \sum_i \left( \frac{k_i}{2} \right) \left( \frac{1}{c_i} + \sum_{j \neq i} \frac{k_j b_{ji}}{2} \right) \]

Where

- \( k_i \): the number of lineages currently in patch i.
- \( c_i \): the proportion of the effective population size of patch i relative to the total effective population size (\( N_i/N \) with \( N_i \) equal to the effective population size of patch i).
- \( h(k_i, k_2, \ldots) \): the rate at which coalescence in patch i occur.
- \( b_{ij}/2 \): is the backward migration rate from patch i to patch j (equal to \( Nb_{ij} \), with \( b_{ij} \) being the backward migration probability from patch i to patch j).

The above equation can then be rewritten as:

\[ h(k_1, k_2, \ldots) = 2N\mu \sum_i \left( \frac{k_i(k_i - 1)}{4N\mu} + \sum_{j \neq i} \frac{k_j b_{ji}}{2\mu} \right) \]

\[ = 2N\mu \sum_i \left( \frac{k_i(k_i - 1)}{\Theta_i} + \sum_{j \neq i} k_i M_{ij} \right) \]

to correspond to the notation presented in (Beerli & Felsenstein 1999, 2001; Beerli 2002), where \( \mu \) is the per-site per generation mutation rate, and \( M_{ij} = b_{ij}/2\mu \). The software migrate estimates \( \Theta_i = 4N\mu \) and \( \gamma = 4N\mu b_{ij}/2 = 4N\mu M_{ij} \) in this case.

The probability of a coalescence event occurring at the end of that waiting time is (also following Nordborg 2001):

\[ \frac{\left( \frac{k_i}{2} \right) / c_i}{h(k_1, k_2, \ldots)} \]

and the probability of a migration event occurring is

\[ \frac{k_i b_{ij}/2}{h(k_1, k_2, \ldots)} \]

Based on this theory, the topology can be built by tracking the coalescence events. The lengths of the branches represent the time between coalescences, which might include a number...
of migration events. Hence, the time between coalescences will be the sum of waiting times until migration and a time until a coalescence event after the last migration.

**MIGRATE**

*MIGRATE* assumes the usual Wright-Fisher model; patches have a constant effective size through time, the rate of mutation is constant, and patches exchange migrants with constant rates per generation (Beerli & Felsenstein 1999, 2001).

Given the genealogy $G$ (topology and branch lengths) and the model of evolution we can calculate the likelihood of a certain dataset $D$ ($Pr(D \mid G)$) (Swofford *et al.* 1996; Huelsenbeck & Crandall 1997; Felsenstein 2003). Given the evolutionary parameters $P$ (the mutation rate, the effective population size, and the migration rates), we can calculate the likelihood of a topology with certain branch lengths ($Pr(G \mid P)$). Summing over all possible genealogies we find the likelihood of the data under a certain set of evolutionary parameters.

\[
L(P) = \sum_G Pr(D \mid G)Pr(G \mid P)
\]

Beerli & Felsenstein (1999, 2001) let the genealogy specify the times and places of the migration event as well as the times of coalescence. Accordingly, they calculate $Pr(G \mid P)$ by finding the products of the probabilities of no event happening in $T$ time intervals associated with a certain genealogy and then multiplying it by the probability of a migration or coalescence happening at the start of these intervals (the bottom of these intervals).

Due to the fact that the genealogy space is infinite, *MIGRATE* uses a Markov Chain Monte Carlo (MCMC) sampling strategy to compute a likelihood ratio

\[
\frac{L(P)}{L(P_o)} \approx \frac{1}{G} \sum_{i=1}^{g} \frac{Pr(G_i \mid P)}{Pr(G_i \mid P_o)}
\]

with $P_o$ representing the parameters used to sample the genealogies $G$, and $g$ is the number of sampled genealogies as described in Beerli & Felsenstein (1999, 2001). The set of initial parameters $P_o$ is introduced using $F_{ST}$ or any other source or method that can provide such initial values.

The search strategy of the genealogical space is described thoroughly in Beerli & Felsenstein (1999). The initial genealogy is generated using a UPGMA method then a minimal number of migration events are added using Sankoff’s parsimony method (Swofford *et al.* 1996; Beerli & Felsenstein 1999; Felsenstein 2003). Time between events are added using an exponential distribution with rate as given in equation 1. A coalescent node or a tip is chosen at random from the current genealogy. The lineage below it is dissolved. This node is then used as a starting point to simulate a coalescent-with-migration process as described above to rebuild that part of the tree until coalescence occurs again (Beerli & Felsenstein 1999). This results in a new genealogy. The process is repeated $g$ times using an accepted previous genealogy in each time to generate the sample of genealogies used in calculating the likelihood ratio. The acceptance rule is based on a Hastings sampling term as described in Beerli & Felsenstein (1999).

**Performance evaluation**

To evaluate the performance of *MIGRATE*, we simulate sequence data as follows. We assume that migration is occurring between two populations with equal effective population sizes as in Beerli & Felsenstein (1999). The migration rate between these two populations is assumed to be symmetric and the mutation rate is taken to be constant for both populations. As in Beerli & Felsenstein (1999), we use two sequence lengths (500 bp and 1000 bp) in our study and assume these data come from a single genetic locus. While one can perform a multi-locus estimate of migration rate leading to a more robust estimate (Brumfield *et al.* 2003) most studies in population genetics across a diversity of organisms are still performed on mtDNA and this has been the dominant use of *MIGRATE* (Rawson *et al.* 2003; Zeh *et al.* 2003). Data used for multilocus analysis has the confounding difficulty of recombination which is typically ignored by multilocus estimators of migration rate and can have a profound impact on these rate estimates (Schierup & Hein 2000).

Genealogies are simulated under the coalescent-with-migration model. The trees are based on a sample of 25 individuals from each population. The simulation uses three different $\Theta_i$ values (0.00025, 0.0025, 0.025) crossed with four different migration rates $4N\mu_i$ (0.0000025, 0.025, 0.25, 2.5) to construct 1000 evolutionary trees per-combination per-sequence-length (500 bp and 1000 bp), where $i$ refers to the population and it is either 1 or 2, $\Theta_i = 4N\mu_i$ and $4N\mu_i = 4N\nu_i/2$. This results in 24,000 trees. These parameter values fall within the range seen in recent studies of migration rates and genetic diversity using *MIGRATE* (Rawson *et al.* 2003; Roman & Palumbi 2003; Zeh *et al.* 2003). The Jukes-Cantor (JC) model of evolution is then applied via SEQ _GEN_ 1.2.5 (Rambaut & Grassly 1996) to the simulated trees to generate the sequence-data that we analyse.

The generated sequences are processed through *MIGRATE* (Beerli 2002) adjusting the transition/transversion ratio to 0.50 to accommodate the JC model. Otherwise, we use *MIGRATE*’s default settings in our runs. Processing is done on a 64-node Beowulf cluster of the University of Idaho.

We assess the results through a number of different criteria. First, we calculate the means, medians, standard errors and coefficients of variation of the estimated parameters to assess the bias of the estimators and their spread. Second, we plot and analyse the sampling distributions of these...
parameters. Finally, we assess the profile confidence intervals outputted by Migrate by counting the number of times they capture the true parameter that we simulate under.

Results

In this section we tackle some of the convergence issues that were associated with using the default settings of Migrate after that we present our results of performance for three different estimators. First we present results on the estimation of genetic diversity or theta, \( \Theta_i \). Next we present results on the performance of estimating migration rates, \( 4N_m_{ij} \), \( \Theta_i \), and \( 4N_m_{ij} \) as described above, and \( i \) is either 1 or 2. Finally, because asymptotic theory is not valid for many complex models used in population genetics, the only valid approach to obtaining confidence intervals is to simulate the distribution of the maximum likelihood estimates. However, enough simulations need to be performed over a wide range of parameter values to get an accurate view of the error structure. Our final results evaluate the confidence intervals associated with estimates of mutation rates and migration rates.

Convergence issues

Running our simulated datasets using the default settings of Migrate generated 79 problematic results with estimates of migration rates and theta well above 1000. These problematic results concentrated in the \( (\Theta_i = 0.00025, 4N_m_{ij} = 0.25) \) \( (\Theta_i = 0.00025, 4N_m_{ij} = 2.5) \), and \( (\Theta_i = 0.0025, 4N_m_{ij} = 2.5) \) parameter combinations of the 500 bp datasets and in the \( (\Theta_i = 0.00025, 4N_m_{ij} = 0.25) \) \( (\Theta_i = 0.00025, 4N_m_{ij} = 2.5) \), and \( (\Theta_i = 0.0025, 4N_m_{ij} = 0.25) \) parameter combinations used to generate data with 1000 bp sequences. Kuhner et al. (2000) argue that there is a small probability that simulations will result in data that might produce an infinite estimate of the \( \Theta \). This might occur when the last two lineages in the total population do not spend enough time in the same patch to coalesce before one of them migrates. This might be the case at reasonably high migration rates. This is expected at any level of mutation, especially the high levels, and not only at the low levels as we see in Tables 1 and 2. Using clustal w (Thompson et al. 1994) as a way to review these data sets we found no justification to remove such sets from our analysis. Detected segregating sites where counted to be between one and 12 for 78 of the datasets. The last dataset had 20 segregating sites corresponding to a 1000 bp sequence length with parameters \( (\Theta_i = 0.0025, 4N_m_{ij} = 0.25) \).

To test whether this is a convergence problem we ran these datasets using 10 short chains with 50 as our increment and 5000 as our sample (chain length = 250 000) and three long chains with increment of 50 and sample of 50 000 (chain length 2 500 000). This resulted in reasonable estimates for 65% of the datasets. The remaining sets where
run using the same short and long chain lengths though we used four chains and a heating scheme of (1, 1.2, 1.5, and 3) to better search the parameter space for the MLEs. We used the Gelman’s R option provided by Migrate (with R = 1.2) to guarantee convergence of the last chains. Only four datasets did not converge after such a rigorous search within two days. Two of these datasets where associated with (θ = 0.0025, 4Nmi = 0.25) parameter combination, and a 500 bp sequence length. The first had six segregating sites and the second had 12. The other two where associated with the parameter combination (θ = 0.000025, 4Nmi = 2.5) and with the 1000 bp sequence length. Both of them had three segregating sites. Although there is still no reason to drop these datasets from our analysis, we nonetheless did with minimal impact on our results. Presumably, with additional computational effort, these data sets would eventually converge as well.

Comparing Tables 1 and 2 we can see that the coefficients of variation and the standard errors become smaller as we increase the length of the sequence.

The medians, on the other hand, indicate that the MLEs overestimate the θi’s only at the lowest level of mutation and migration. Otherwise the MLEs tend to underestimate the θi when its true value is small. As the true parameter increases the median becomes closer to the true value. The median does best at the true value, θi = 0.025. The medians come closer to the real value as the migration rate increases. These medians are positively biased at the highest migration rates. The medians tend to be closer to the truth as we increase the sequence length. This can be seen for both θi = 0.0025 and θi = 0.025 in Table 2 where the medians seem to estimate the truth very well.

Figures 1–3 show the sampling distributions under the θi-migration-rate combinations for the first population for the 500 bp sequence data. The graphs look the same for the second population. The heavy tail of the distribution is quite obvious when the true θi equals 0.00025. The so-called ‘fatal attraction to the zero’ (Beerli & Felsenstein 1999) is also clear in this case where the MLEs cluster near zero especially in the case of the two smallest migration rates. As θi increases the distribution tends to the bell-shape, though still right-skewed. A vertical axis is introduced into the graphs to show the location of the true parameter. The ‘fatal attraction to zero’ seems to reduce drastically as θi and the migration rate increase. It disappears at the highest level of θi. Figures 4–6 show the sampling distributions

---

**Table 2** Summary statistics of MLEs of Θ for DNA sequence length = 1000 bp

<table>
<thead>
<tr>
<th>True Θ1</th>
<th>0.00025</th>
<th>0.0025</th>
<th>0.025</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Estimates</td>
<td>True 4Nmi</td>
<td>0.00072</td>
<td>0.00029</td>
</tr>
<tr>
<td>Median of Estimates</td>
<td>True 4Nmi</td>
<td>0.00027</td>
<td>0.00018</td>
</tr>
<tr>
<td>SE of Estimates</td>
<td>True 4Nmi</td>
<td>0.00012</td>
<td>0.00036</td>
</tr>
<tr>
<td>Cv of Estimates</td>
<td>True 4Nmi</td>
<td>1.60</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Comparing Tables 1 and 2 we can see that the coefficients of variation and the standard errors become smaller as we increase the length of the sequence.

The medians, on the other hand, indicate that the MLEs overestimate the θi’s only at the lowest level of mutation and migration. Otherwise the MLEs tend to underestimate the θi when its true value is small. As the true parameter increases the median becomes closer to the true value. The median does best at the true value, θi = 0.025. The medians come closer to the real value as the migration rate increases. These medians are positively biased at the highest migration rates. The medians tend to be closer to the truth as we increase the sequence length. This can be seen for both θi = 0.0025 and θi = 0.025 in Table 2 where the medians seem to estimate the truth very well.

Figures 1–3 show the sampling distributions under the θi-migration-rate combinations for the first population for the 500 bp sequence data. The graphs look the same for the second population. The heavy tail of the distribution is quite obvious when the true θi equals 0.00025. The so-called ‘fatal attraction to the zero’ (Beerli & Felsenstein 1999) is also clear in this case where the MLEs cluster near zero especially in the case of the two smallest migration rates. As θi increases the distribution tends to the bell-shape, though still right-skewed. A vertical axis is introduced into the graphs to show the location of the true parameter. The ‘fatal attraction to zero’ seems to reduce drastically as θi and the migration rate increase. It disappears at the highest level of θi. Figures 4–6 show the sampling distributions
under the 1000 bp sequence data. These graphs show the same patterns as the 500 bp graphs. A slightly lower variation is noticed at all levels and overcomes the ‘fatal attraction’ phenomenon at the 0.0025 $\Theta$-level.

Figure 7 introduces box-plots of the means and medians of the $\Theta$'s for the 500 bp and 1000 bp sequence lengths, respectively, for the first population. The tendency of the mean to overestimate is quite clear from these graphs. The medians, on the other hand, perform much better with less variation and less bias.

**Migration rate estimation**

Table 3 introduces the results for the migration rates. At the two lowest migration levels (0.0000025 and 0.025) the means decrease going from $\Theta = 0.00025$ to $\Theta = 0.0025$ and then increase again. These means are positively biased. At the two high levels of migration the means always decrease, overestimating all way through. The standard errors tend to decrease as $\Theta$ increases and increase as the migration rates increase. The coefficients of variation do not follow a clear
trend except when $\Theta_i = 0.025$ where they seem to decrease as the migration rate increases. It is worth noting that these coefficients of variation are always greater than one, highlighting the high variation present in estimating the migration rate.

As the $\Theta_i$ increases, the medians increase. Regardless of the lowest migration level where the medians always overestimate the true migration rate — the medians almost always underestimate the true values. This underestimation is most noticeable at the lowest level of $\Theta_i$.

Table 3 introduces the results for the 1000 bp sequence length and shows that the means overestimate the migration rates drastically still; even with the removal of the problematic datasets indicated in the Convergence issues subsection. This overestimation is the result of the large variation in the obtained estimates under the current setup. This high variability is clearly seen in the standard errors and the coefficients of variation.

Figures 8–13 introduce the sampling distribution of the migration rates under the different levels of $\Theta_i$ and the two
sequence lengths for the first population. Similar graphs result for the second populations (results not shown). Both the fatal attraction to zero and the extremely heavy tail of the MLEs are quite obvious. The distribution has a very large variation. Only at migration rates of 0.025 and 0.25 do the sampling distributions tend to look, very remotely, bell-shaped, as the $\Theta_j$ increases and the sequence length increases. At the highest level of migration the sampling distribution tends always to a J shape.

Figure 14 shows box-plots of the MLE means and medians for the 500 bp and 1000 bp sequence lengths for the first population. The tendency of the means to overestimate at times is quite clear. Also clear is the medians consistent underestimation of the true migration rates. The variation in both means and medians increase dramatically as the migration rates increase. This reflects the performance of the estimates under the different levels of $\Theta_j$.

### Confidence intervals

MIGRATE introduces profile likelihood confidence intervals (CI’s) (Meeker & Escobar 1995) for the estimated parameters.

TABLE 4 Summary statistics of MLEs of the migration rate for DNA sequence length = 1000 bp

<table>
<thead>
<tr>
<th>True $4N_i m_{12}$</th>
<th>2.50E-06</th>
<th>0.025</th>
<th>0.25</th>
<th>2.5</th>
<th>True $4N_i m_{21}$</th>
<th>2.50E-06</th>
<th>0.025</th>
<th>0.25</th>
<th>2.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00025</td>
<td>0.0392</td>
<td>0.27</td>
<td>3.81</td>
<td>9.38</td>
<td>0.00025</td>
<td>0.042</td>
<td>0.280</td>
<td>2.47</td>
<td>11.48</td>
</tr>
<tr>
<td>0.0025</td>
<td>0.0044</td>
<td>0.058</td>
<td>0.32</td>
<td>3.92</td>
<td>0.0025</td>
<td>0.004</td>
<td>0.058</td>
<td>0.37</td>
<td>3.61</td>
</tr>
<tr>
<td>0.025</td>
<td>0.0182</td>
<td>0.069</td>
<td>0.31</td>
<td>2.65</td>
<td>0.025</td>
<td>0.018</td>
<td>0.077</td>
<td>0.32</td>
<td>2.76</td>
</tr>
<tr>
<td>Median of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Median of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00025</td>
<td>4.00E-06</td>
<td>0.004</td>
<td>0.12</td>
<td>1.16</td>
<td>0.00025</td>
<td>3.00E-06</td>
<td>0.003</td>
<td>0.13</td>
<td>1.20</td>
</tr>
<tr>
<td>0.0025</td>
<td>2.55E-05</td>
<td>0.019</td>
<td>0.17</td>
<td>1.90</td>
<td>0.0025</td>
<td>2.80E-05</td>
<td>0.022</td>
<td>0.19</td>
<td>1.89</td>
</tr>
<tr>
<td>0.025</td>
<td>5.94E-04</td>
<td>0.031</td>
<td>0.21</td>
<td>1.92</td>
<td>0.025</td>
<td>5.90E-04</td>
<td>0.036</td>
<td>0.22</td>
<td>2.11</td>
</tr>
<tr>
<td>SE of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>SE of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00025</td>
<td>0.100</td>
<td>2.31</td>
<td>25.63</td>
<td>36.01</td>
<td>0.00025</td>
<td>0.103</td>
<td>3.81</td>
<td>13.77</td>
<td>52.15</td>
</tr>
<tr>
<td>0.0025</td>
<td>0.017</td>
<td>0.108</td>
<td>0.59</td>
<td>8.43</td>
<td>0.0025</td>
<td>0.012</td>
<td>0.103</td>
<td>0.85</td>
<td>6.47</td>
</tr>
<tr>
<td>0.025</td>
<td>0.043</td>
<td>0.117</td>
<td>0.41</td>
<td>2.80</td>
<td>0.025</td>
<td>0.046</td>
<td>0.141</td>
<td>0.43</td>
<td>2.57</td>
</tr>
<tr>
<td>CV of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CV of Estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00025</td>
<td>2.55</td>
<td>8.51</td>
<td>6.73</td>
<td>3.84</td>
<td>0.00025</td>
<td>2.47</td>
<td>13.60</td>
<td>5.58</td>
<td>4.54</td>
</tr>
<tr>
<td>0.0025</td>
<td>3.80</td>
<td>1.84</td>
<td>1.83</td>
<td>2.15</td>
<td>0.0025</td>
<td>3.24</td>
<td>1.78</td>
<td>2.28</td>
<td>1.79</td>
</tr>
<tr>
<td>0.025</td>
<td>2.36</td>
<td>1.69</td>
<td>1.33</td>
<td>1.06</td>
<td>0.025</td>
<td>2.54</td>
<td>1.83</td>
<td>1.32</td>
<td>0.93</td>
</tr>
</tbody>
</table>

© 2004 Blackwell Publishing Ltd, Molecular Ecology, 13, 837–851
Table 5 shows the percentage of times these 95% CI's managed to capture the true parameter under which the data was simulated.

For the $\Theta_i$'s, the confidence intervals did well at the highest level (0.025). They slightly improved as the sequence length increased. These proportions (at $\Theta_i = 0.025$) reduced as the migration rates increased. In contrast, at the 0.0025 $\Theta_i$-level the proportions where much the same, except when the migration rate went up to 2.5. At this point the proportions decreased drastically in both the 500 bp and 1000 bp cases. The length of the sequence made a good difference in the capture proportions; as the length increased there was quite an increase in the proportion of the CI's that captured the true $\Theta_i$.

At the lowest level of $\Theta_i$ (0.00025) the proportions did very poorly increasing as the migration rates increased.
Again we note a good improvement in the capture proportions as the sequence length increases.

Regarding the migration rates, the proportions captured by the CI’s were highest in two cases. First at $\Theta_i$ level 0.025 and migration rates 0.25 and 2.5 where the CI’s captured the true parameter 60% to 77% of the times (there was a slight improvement between the two sequence lengths). The proportions of times the true migration rates were captured were lowest at the two lowest true migration levels. The second case corresponds to $\Theta_i = 0.0025$ and migration rates of 0.0000025 and 0.25 with proportions ranging between 58% and 65%. The proportions did not improve much as the sequence length increased. These proportions reduced as migration rates increased. At the lowest level of $\Theta_i$, the proportions were poor at all levels at the 500 bp sequence length (with a maximum of 48.3%). Not much improvement occurred when the length increased. As the migration rates increased the capture proportions reduced in general.

Discussion

The results indicate that MIGRATE estimates $\Theta_i$’s better than it does the migration rates. MLEs of the $\Theta_i$’s tend to the
normal distribution as the sequence length increases. It might be that the sequence lengths and the number of loci used in our analysis are not enough for \texttt{migrate} to be able to detect migration appropriately. This might be due to the tremendous right skewness of the sampling distribution of the migration rates. Yet these are typical values for mtDNA studies that use \texttt{migrate} to estimate migration (Roman & Palumbi 2003).

\textit{The thetas}

The length of the sequence and the parameter combination significantly impact the confidence associated with the estimates of the $\Theta_i$'s. As the sequence length increases, we are more confident in our estimates. This is reflected in the improved performance of the confidence intervals with the increase in the sequence length as indicated in the results.
Moreover, as the true \( \Theta_i \) increases we are more confident in our estimates of it at low levels of migration. This indicates that \textsc{migrate} will perform relatively well in estimating \( \Theta_i \) with data of higher divergence such as mitochondrial sequence data compared to nuclear data. It also indicates that migration will impact the estimate of the \( \Theta_i \); as migration-rate increases the estimates of the \( \Theta_i \)'s will be biased upward. The best estimates of the \( \Theta_i \)'s occur when migration is low and the true \( \Theta_i \) is high with long sequence data.

The variation in the estimates decreases as the length of the sequence increase and as the true \( \Theta_i \) increases at low migration levels. This is reflected in the behaviour of the coefficient of variation seen in the results.

\textit{Migration Rate}

\textsc{migrate} did not accurately estimate migration rate. While this may be primarily due to the inherently small amount of signal in the data, our largest concern is with the use of
profile likelihoods to assess the variability in the estimates. Profile likelihood methods considerably under represent the error in the migration rate estimates. The sampling distribution of the migration rates is heavy-tailed right-skewed indicating that \textsc{migrate} tends to always underestimate this parameter with a high probability of overestimating occasionally. The medians of the sampling distribution show the underestimation tendency of the MLEs, while the means reflect the effect of the occasional overestimation problem. The estimates tend to improve as $\Theta_i$ increases at the end of the range of the true migration rates (the 0.25 and 2.5). The variation in the estimates is high which is clear from the resulting, (always greater than one), coefficients of variation. Not much improvement in variation occurs with the increase of the length of the sequence. However, good performance of the migration estimates was detected in respect to datasets generated under true migration parameters of 0.25 and true $\Theta_i$ equal to 0.025.

One should be cautious in using the current available version of \textsc{migrate} in estimating migration rates under similar settings as presented in our simulations. Also, one should not use the profile confidence intervals generated by \textsc{migrate} for such estimates.
Acknowledgements

We thank Laurent Excoffier for the invitation to submit to this special issue and for his patient and careful handling of our manuscript. We thank two anonymous reviewers for their helpful comments to improve our manuscript. We also thank Peter Beerli for his contribution and important remarks. This Research is partially sponsored by the Initiative in Bioinformatics and Evolutionary Studies (IBEST) at the University of Idaho. Funding was provided by NSF EPSCoR, EPS-0080935, NSF EPSCoR, EPS-0132626, and NIH NCRR grant NIH NCRR-1P20RR16448-01. Paul Joyce is also funded by NSF DEB-0089756, and NSF DMS-0072198 and Keith Crandall by NIH R01AI50217 and R01GM66276.

References


Fig. 14 Boxplots of the Means and Medians of the MLRs of Migration-Rate’s Sampling Dist. for the First Population.


