Genetics: Published Articles Ahead of Print, published on February 16, 2005 as 10.1534/genetics.104.034199

Bayesian estimation of recent migration rates after a spatial expansion

Grant Hamilton[1], Mathias Currat[1, 2], Nicolas Ray[1, 2], Gerald Heckel[1], Mark Beaumont[3] and Laurent Excoffier[1]

[1]Computational and Molecular Population Genetics Lab, Zoological Institute, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[2]Genetics and Biometry Laboratory, Dept. of Anthropology and Ecology, University of Geneva, CP 511, 1211 Geneva 24, Switzerland

[3]School of Animal and Microbial Sciences, The University of Reading, Whiteknights, Reading RG6 6AJ, United Kingdom.

Running title: Bayesian estimation after a spatial expansion

Corresponding author:

Laurent Excoffier,

Computational and Molecular Population Genetics Lab,

Zoological Institute, University of Bern,

Baltzerstrasse 6,

3012 Bern, Switzerland

email:laurent.excoffier@zoo.unibe.ch

ABSTRACT

Approximate Bayesian Computation (ABC) is a highly flexible technique that allows the estimation of parameters under demographic models that are too complex to be handled by full likelihood methods. We assess the utility of this method to estimate the parameters of range expansion in a 2 dimensional stepping stone model, using samples from either a single deme or from multiple demes. A minor modification to the ABC procedure is introduced which leads to an improvement in the accuracy of estimation. The method is then used to estimate the expansion time and migration rates for five natural common vole populations in Switzerland typed for a sex linked marker and a nuclear marker. Estimates based on both markers suggest that expansion occurred less than 10 kyr ago, after the most recent glaciation, and that migration rates are strongly male biased.

INTRODUCTION

Making quantitative inferences on molecular data in complex demographic settings remains an ongoing challenge. Traditionally, inferences have been made using summary statistics under simplified models (e.g. FU and CHAKRABORTY 1998). While sometimes useful for qualitative comparisons, these simple models do not adequately reflect the complexity of processes which might impact on molecular genetic diversity. Recent advances in maximum likelihood and Bayesian approaches have shown that it is possible to make full use of the data gathered from population samples (e.g. BEAUMONT 1999; BEERLI and FELSENSTEIN 2001; NIELSEN and WAKELEY 2001; WANG and WHITLOCK 2003). Although these methods have the potential to be accurate, the calculation of likelihoods under complex models can be problematic, necessitating the use of simplified demographic models. A promising alternative approach is to compare summary statistics that are calculated from observed data and related to the parameter(s) of interest, with summary statistics simulated under a model for which the parameters are known (e.g. TAVARÉ *et al.* 1997; FU and LI 1997; PRITCHARD *et al*. 1999; ESTOUP *et al*. 2001; ESTOUP and CLEGG 2003; ESTOUP *et al*. 2004). An appealing feature of these approximate Bayesian computational (ABC) methods is that models of any complexity can be used, provided only that data can be simulated under the model (BEAUMONT *et al*. 2002). This allows for the use of models that more closely reflect the complexity of real processes, potentially allowing the estimation of more meaningful biological parameters.

Many species have had a complex history that has included a spatial expansion from a restricted range (e.g. an expansionary period following an ice age), with the establishment of new demes and the exchange of genes among those demes (HEWITT

2000; RAY *et al*. 2003; EXCOFFIER 2004). Although these expansion processes affect various aspects of molecular diversity differently, and can thus be described using a combination of summary statistics, making joint estimates of expansion parameters in an explicitly spatial setting would be difficult using either full likelihood or more conventional methods. However, it may be possible to use an ABC approach to simultaneously infer spatial expansion parameters in the context of an appropriate spatial model. Additionally, data from differently inherited molecular markers can be used to investigate interesting biological phenomena such as differences in dispersal rates between the sexes after a range expansion.

The purpose of this study is to show that an approximate Bayesian approach can be used to accurately infer parameters of spatial expansion in a two dimensional stepping stone (2DSS) model, using both mtDNA sequences and Short Tandem Repeats (STRs) as molecular markers. We test the method using a range of parameter combinations and also show that sampling from several demes rather than a single deme improves estimation accuracy. Finally, we challenge the method with empirical data from natural populations of the common vole (*Microtus arvalis*) to assess differences between male and female dispersal levels, and to determine the time since the most recent range expansion.

MATERIAL AND METHODS

Estimation procedure and simulation model

We follow the approximate Bayesian computation (ABC) approach described formally by BEAUMONT *et al*. (2002), and briefly below. The method relies on the simulation of large numbers of datasets using known parameters under a given model. Summary statistics are calculated for each data set. These simulated statistics are then

compared with summary statistics calculated from the observed sample. Simulated summary statistics that are 'close' enough (i.e. fall within a pre-determined distance, δ) to the observed summary statistics are retained, with their associated parameters. All other data sets are rejected. After a smooth weighting and local regression adjustment step, the accepted parameters form an approximate posterior distribution that can be used to estimate the parameter of interest.

Modification to conventional ABC: One aspect of simultaneously estimating several parameters with the ABC method is that while a statistic may estimate a given parameter well, it may be a poor estimator for other parameters. This problem has the potential to lead to a decrease in the accuracy of estimation. For instance, while variance in the size of STR alleles evolving under a stepwise mutation model shows a strong relationship with the scaled expansion time (τ) (GOLDSTEIN *et al*. 1999), it carries little information on gene flow. Conversely, it would be expected that the scaled migration rate $M=2Nm$ (where N is the number of genes present in a local deme and m is the immigration rate from neighbouring demes) would show a stronger relationship with F_{ST} than, for example, the mean number of pairwise differences among mtDNA sequences (EXCOFFIER 2004). Thus while statistics are chosen because they describe well one aspect of a range expansion phenomenon, they will usually be less informative on other parameters. We address this issue by using a weighting scheme such that statistics carrying more information on the parameter of interest are given a greater weight. BEAUMONT *et al.* (2002) defined the distance δ using an Euclidean metric to give circular acceptance regions, accepting a proportion Pδ of the simulation data (the tolerance). To aid estimation efficiency, we use a Weighted Euclidean Distance (WED). We assess the relationship between each parameter-statistic pair during a preliminary analysis step, in which a local regression

is run in the vicinity of each observed summary statistic. The square of the correlation coefficient between a parameter and a statistic (R^2) from each local regression was taken as a simple measure of the utility of the statistic to act as an estimator of the parameter in that region of the distribution.

For illustration, consider just two statistics, S_1 and S_2, that are believed to be informative for the parameter Φ_1, with values s_1 and s_2 calculated for an observed data set (this procedure could of course be extended to more than 2 statistics). To calculate the weight for S_1 for this observed data set, simulated statistics that are in the vicinity of s_1 (for example, the 1% of the empirical distribution closest to s_1) are collected, together with their associated parameters. The retained statistics are regressed on the retained parameters, allowing the weight to be calculated as

$$w_{ij} = -\log(1 - R_{ij}^2), \quad (1)$$

where R_{ij}^2 is the determination coefficient of the *i*-th parameter by the *j*-th statistic. After the operation is repeated for S_2, the weights are scaled so as to sum to unity as $w_{ij}^* = w_{ij} / (w_{i1} + w_{i2}),\ j = 1, 2$. The weighted Euclidean distance δ_i between the "observed" values s_1 and s_2 and the simulated values of the same statistics s_1' and s_2' can then simply be found as

$$\delta_i = \sqrt{w_{i1}^* (s_1 - s_1')^2 + w_{i2}^* (s_2 - s_2')^2}, \quad (2)$$

where the index *i* emphasizes the fact that the Euclidean distances between simulated and observed statistics will differ depending on the parameter to be estimated. Note that a weighting scheme simply using R^2 was also investigated, but did not lead to drastic differences to the present weighting scheme. Compared to R^2, an advantage of the weighting in eq. 1 is that it will enhance the difference in weight between parameter-statistic pairs with a strong relationship (R^2>0.8) and those with a poor

relationship, ensuring that most weight is placed on statistics with large determination coefficients. After the weighting step, each parameter is independently estimated in a way similar to that of the standard ABC multivariate local regression approach. Apart from the use of a WED, our estimation procedure differs from BEAUMONT *et al.* (2002) only in that we separate the time consuming simulation stage from the estimation procedure. The modified algorithm is: (1) select prior distributions for parameters Φ_1 and Φ_2; (2) draw values Φ_1' and Φ_2' for Φ_1 and Φ_2 from the appropriate priors; (3) use Φ_1' and Φ_2' to simulate genetic diversity under the chosen model to produce a single dataset, using the same genetic marker, number of samples and number of loci as the observed data; (4) calculate the summary statistics values s_1' and s_2' for the dataset; (5) repeat steps 2-4 until the desired number of simulations have been completed; (6) compute summary statistics s_1 and s_2 on observed data, (7) choose a tolerance level P_δ to determine the number of data points retained for posterior density estimation; (8) for S_1, retain a small proportion (e.g. 1%) of the s_1' that are closest to s_1 together with their associated parameters Φ_1'; (9) regress the retained s_1' on Φ_1' and obtain the R^2 value; (10) repeat steps 8 and 9 for S_2 and calculate weights $w^*_{\Phi_1 S_1}$ and $w^*_{\Phi_1 S_2}$ using Equation 1 and scaling described above; (11) calculate the WEDs between each simulated data set using $w^*_{\Phi_1 S_1}$ and $w^*_{\Phi_1 S_2}$ and the observed data as in eq. 2, and retain a fraction P_δ of the data sets (summary statistics with associated parameters) that are closest to the observed data; (12) perform multiple regression of the retained statistics on the parameters and adjust posterior density as described in BEAUMONT *et al.* (2002); (13) repeat steps 8-12 for Φ_2.

The difference in analysis time between the simultaneous estimation of parameters under the standard ABC method, and the extended method described above in which parameters are estimated in succession is small, being in the order of seconds to several minutes for a single set of observed data, depending on the size of the simulation file. Also the time necessary for computing the weights is negligible compared to the time required for the simulations and the rest of the estimation procedure. A comparison between the standard ABC method and the extended method using a WED was conducted using the demographic model, summary statistics and priors that are described below.

Simulations: The simulation model is a modification of the model described by RAY *et al*. (2003). A subdivided population was simulated on a torus of 50 by 50 demes with an expansion from an originating deme arbitrarily located at <25,25>. RAY *et al*. (2003) simulated a forward demographic expansion in which colonisation occurs as a wave from the originating deme, with an ongoing exchange of migrants among demes that have been colonised (range expansion (RE) model). Our simulation model differs from the RE model, making the common assumption of an instantaneous expansion, such that each of the 2500 demes is filled instantly to a given size N (note that N represents here the number of genes present in a deme) identical for all demes (instantaneous expansion (IE) model). This slight modification allowed for much faster simulations. Under this model, neighbouring demes exchange genes at a rate m during the T generations following the expansion and a standard backward coalescent process is implemented to simulate gene genealogies. During this backward process, gene lineages can either migrate between different demes or coalesce if they are in the same deme. Going backward in time, the instantaneous expansion corresponds to an instantaneous contraction, where all gene lineages are instantaneously brought back to

the originating deme, where further coalescent events can occur until only one lineage remains. Genetic data are obtained by adding mutations at rate μ under a strict stepwise mutation model for STRs, and a finite sites model without transition bias for sequences. Parameters describing the scaled expansion time $\tau = 2T\mu$, the scaled population size $\theta = 2N\mu$ and the scaled migration rate $M = 2Nm$ are thus known and are recorded for each simulation. We note here the advantage of separating the simulation model from the estimation procedure. Since the accuracy of estimation depends in part on the number of simulations used (the simulation size), where possible it is desirable to use several hundreds of thousands to millions of simulations. Depending on the complexity of the underlying demographic model, it therefore takes much longer to generate the simulation file than to run the ABC estimation procedure. During the exploration phase of research, the separation of these two steps saves considerable time by allowing multiple analyses to be run using a single simulation file.

Samples and summary statistics: Samples were drawn from three demes on the torus, which were located arbitrarily at <10,40>, <30, 30> and <40, 15> (identified hereafter as demes 1, 2 and 3 respectively). There were at least 10 demes between each of the sampling demes. Fifty genes were drawn from each sampled deme, consisting of either DNA sequences (300 bp) or nuclear STRs (10 independent loci), which were used to calculate summary statistics. For mtDNA, four within-deme summary statistics were calculated for each sampled deme: number of haplotypes, k; homozygosity, H_o; number of segregating sites, S; and the average number of pairwise differences, π. For STR data, three within-deme summary statistics were calculated for each sampled deme: mean number of alleles per locus, a; homozygosity, H_o; and mean variance (across loci) in allele repeat number. The fixation index, F_{ST}, was

calculated among demes for both markers whenever more than one deme was sampled. These statistics were chosen for their ability to estimate parameters based on preliminary investigations and their known dependency on the values of the parameters of a range expansion under the infinite island model (EXCOFFIER 2004).

Parameter estimation: For mtDNA, simulations were sampled from prior distributions of: τ as a uniform distribution between 0 and 50, θ as a log-uniform distribution between 0.01 and 10, and M as a log-uniform distribution between 0.01 and 500. For STR loci, simulations were sampled from prior distributions of: τ as a uniform distribution between 0 and 300, θ as a log-uniform distribution between 0.01 and 200, and M as a log-uniform distribution between 0.01 and 500. A flat uniform distribution was used for τ because there is often no prior expectation for the time of the range expansion. However, for M and θ, we used log-uniform prior distributions, in order to have equal coverage (and potentially equal accuracy of estimates) for small (<1), intermediate (>1 and <10) and large (>10) values of the parameters. Unless otherwise stated, 1 000 000 values of the summary statistics were generated and a tolerance $P_\delta = 0.001$ was used to give 1000 points from which parameters were estimated. As suggested by BEAUMONT *et al*. (2002), a log transformation was applied to the retained, simulated parameters before the regression adjustment, and parameter estimates were based on the back transformed values. We also follow BEAUMONT *et al*. (2002) (equation 7) in using the fitted value of the regression line as a point estimate of the parameter. During exploratory analysis of a small subset of results, means, medians and modes of the posterior distribution were calculated as alternative estimators, and only minor differences in the value of parameter estimates were found.

RESULTS

Simulation studies

Modified vs. standard ABC: The accuracy of estimation under the ABC procedure detailed by BEAUMONT *et al*. (2002) depends on several factors, including the simulation size, the number of parameters to estimate and the tolerance. For a given simulation size, increasing the number of parameters requires an increase in the tolerance. To determine if using a WED increased estimation efficiency across a range of tolerance values, compared with a standard ABC approach under the 2DSS model, the two methods were assessed as follows. An observation set of sequences was generated under parameters τ= 10, θ = 5 and M=10. Here, and in subsequent evaluations, each observation set consisted of 1000 observations. The parameter chosen for evaluation (M) was estimated for each of the data sets across a range of tolerance levels using both the standard ABC approach (BEAUMONT *et al*. 2002) and an ABC procedure using the WED modification. The relative mean square error (RMSE) was used to assess both bias and accuracy. The WED modification improves estimation efficiency, particularly as the tolerance increases (Figure 1). A similar result was found for STR loci (data not shown). Similar trends were found for the estimation of $\tau$and θ for both markers, although the magnitude of the improvements was smaller (data not shown). In the limit of increasing numbers of simulated data sets, the advantage gained by estimating parameters separately and using a WED decreases until the estimates converge. Indeed, this was previously shown by BEAUMONT *et al*. (2002) when comparing results of the standard ABC method with an earlier rejection algorithm. Although one million simulations is adequate for the estimation of three parameters in the model presented here, the capacity to use such a large number of simulations will depend to some extent on the complexity of the

underlying model and hence simulation times. For more complex models it may be necessary to estimate parameters using fewer simulations. Previous investigations have shown that with fewer (50 000) simulations under the present model, the trend is similar to that shown in Figure 1 as tolerance increased, but the WED approach is considerably more accurate than the standard method. For example, using the mtDNA observation set used above, and under identical conditions except for a simulation size of 50 000, the RMSE for M using $P_\delta = 0.01$ for the standard ABC approach is 0.45 (± 0.028 S.E.), which is substantially greater than that from the WED modification (RMSE = 0.33 ± 0.018 S.E.). The greatest benefit of using the WED approach thus accrues when the number of simulations is small with respect to the number of parameters to be estimated. All further estimation in this study was conducted using the WED modification.

[Figure 1 about here]

The IE model was used to generate observation sets generated under a range of known parameter values of τ, θ and M. Analyses were conducted using samples from a single deme (deme 1), or from the 3 demes described above. For each parameter combination, the behaviour of the estimator was assessed using the mean, and the 2.5 and 97.5 quantiles of the distribution of the 1000 estimates, as well as the number of times the 95% credible intervals of the posterior distributions contained the true parameter value (coverage property).

Parameter estimation from a single deme: When samples are drawn from a single deme, F_{ST} cannot be calculated, and it is therefore not used as a summary statistic. As reported in Table 1a for simulated DNA sequence data, there is a substantial bias in estimates of τ when expansion time is short and the scaled migration rate M is low to moderate. Estimates are more accurate under short expansion times when M

increases, and the range of estimates decreased as shown by the 2.5 and 97.5 quantiles. Migration rate is underestimated when expansion time is short and population size (θ) is small. Again, τ and M are estimated quite well when migration is large, and while the range of estimates for τ is good, 2.5 and 97.5 quantiles show that the distribution of estimates for M is still very broad. Population size estimates were good across the set of parameter combinations with an acceptable range. The coverage shows that reconstructed 95% credible intervals were conservative for most combinations of parameters. Parameter estimates using STR data are better than those obtained using single-locus DNA sequences, as would be expected when using multiple loci for estimates (Table 1b). Although τ is often estimated well, there is again a tendency to overestimation when migration is low to moderate, which is most marked when population sizes are small. Migration estimates are good when M is very small, with an underestimation of the parameter as M increases. The distribution of estimates is however broad. Population size is usually underestimated, although when expansion time is long and M is large, this parameter is overestimated. Coverage study shows that the credible intervals tend to be conservative across most parameter combinations.

Parameter estimation from three demes: There are substantial gains in accuracy across a wide range of parameter value combinations when samples are drawn from 3 demes rather than a single deme (Table 2). For mtDNA sequences, expansion times are consistently estimated well and the range of estimates is relatively narrow (Table 2a). The reduced coverage also implies that credible intervals of the posterior distributions are narrower than for 1 deme estimate while still encompassing the true parameter value in approximately 95% of cases. Migration estimates are also very accurate across the range of parameter combinations simulated. The range of

estimates for *M* is relatively narrow and credible intervals show good coverage properties of the values under which the data sets were generated. Similarly, θ estimates are good with conservative credible intervals. Parameters are also estimated well when using STR data (Table 2b). Expansion time estimates are consistently accurate, with a narrow range of estimates. *M* estimates are now consistently good across the range of parameters simulated, although with a tendency towards slight underestimation when *M* is moderate. The 2.5 and 97.5 quantiles show that the range of estimates of *M* is acceptable, and credible intervals of the posterior distribution show good coverage of the true migration value with little substantial deviation from expectation. Population size is also estimated well and does not show the tendency towards underestimation observed when data are sampled from a single deme. However, the range of estimates of θ is broad for some parameter combinations when compared with the estimate range of other parameters, and credible intervals are conservative.

Application to the estimation of sex-biased dispersal in the common vole (*Microtus arvalis*): The present distributional range of *M. arvalis* spans most of Europe and the western parts of Asia (MITCHELL-JONES *et al.* 1999) and Switzerland in the centre of the Alps has been recently colonized after the melting of the ice cover around 10,000 years ago. The mode of colonization and dispersal in small rodents suggests that the simple 2DSS world is an appropriate spatial model since physical restrictions and social association in subterraneous burrows effectively limit long range dispersal. Dispersal is generally biased towards males in *Microtus* like in the majority of mammals (AARS & IMS 2000; CLOBERT *et al*. 2001; CLUTTON-BROCK 1989), but its effectiveness and the extent of differences between the sexes are unknown since no genetic studies have been conducted. We address these questions

for *M. arvalis* by analyzing a set of five populations (124 individuals) throughout Switzerland for which data from twelve STR loci were typed and a 321 base pair fragment of the hypervariable region II of the mitochondrial control region were available. Detailed information on sampling procedures, molecular methods, and the data set can be found in HECKEL *et al.* (submitted). The five Swiss samples reported in this study have been chosen because they present the same mtDNA ("Central") cytb lineage, which is found in Southern Germany and Northern Switzerland. We therefore assumed that the five Swiss samples belonged to the same expansion wave which has colonized Switzerland from the North after the last glaciation (FINK *et al.* 2004). Simulations were conducted in the 2DSS described previously. The five sampling demes were spread randomly across the homogenous world, subject to the constraint that at least five non–sampled demes lay between each of the sampling demes. Although sampling demes were assigned to fixed positions during simulations, previous investigations have shown that the position of sampling sites has no discernible effect on estimated parameter values. For this, we created 12 simulation files (four replicates, with each replicate consisting of three simulation files). Within each replicate, sampling deme locations were fixed at randomly chosen positions, however these positions differed among replicates. All simulation files were challenged with the same observed data and estimates of τ and M were made to assess whether the variation in estimates among replicates (due to sampling deme position) was greater than variation within replicates. No such difference was found (data not shown).

One million simulations were conducted for each type of molecular marker. Summary statistics for mtDNA and STRs were identical to those used in previous evaluation simulations. For mtDNA, simulations were sampled from prior distributions of: τ

uniform [0:20], θ log uniform [0.01:20] and M log uniform [0.01:500]. For STRs, simulations were sampled from prior distributions of: τ uniform [0:200], θ log uniform [0.01:100] and M log uniform [0.01:500]. The ABC procedure was conducted as described previously. A tolerance of 0.001 was used for each estimation procedure to give 1000 points for the calculation of parameter estimates and the 95% credible intervals of posterior distributions. The inferred scaled expansion time for STRs is 23.0 (Table 3). Using a STR mutation rate of 5×10^{-4} (JARNE and LAGODA 1996, ELLEGREN 2004) and the reasonable assumption of three generations per year (HAUSSER 1995), the time of the most recent range expansion was calculated as approximately 7667 years (with a 95% credible interval of 5455 to 11 337 years). For mtDNA, τ was estimated as 4.8 (Table 3). For the stretch of the control region used in this study, a mutation rate of 46 % per site per million years of divergence was estimated by comparison with a related species with known divergence time (FINK *et al.* 2004), although this estimate may be conservative due to the potential for multiple hits per site. By again assuming 3 generations per year, the inferred expansion time for mtDNA is approximately 8127 years (with a 95% credible interval of 4725 to 13 605 years), showing that the two types of markers are in excellent agreement in showing a post-glacial colonization time of Switzerland. As expected, estimation of the M parameter differs significantly between mtDNA and STR data (Table 3), with M_{STR}=13.2 (95% credible interval [5.6, 20.4]) for nuclear genes and M_{mt}=0.29 [0.02, 2.5] for mtDNA. The difference in ploidy and transmission pattern between mtDNA and STR markers implies that $M_{STR} = 4 N_e m$, where N_e is the effective population size and $N_e m$ is the effective number of nuclear genes (3.3) exchanged between neighboring demes per generation, while $M_{mt} = 2 N_f m$, where N_f is the number of females and $N_f m$ is the number of female genes (0.145) exchanged between demes

per generation. The posterior distribution of the *Nm* values for female, male, and total immigrant genes is shown in Figure 2. The male posterior distribution was obtained from the convolution of the total and female densities, under the simple assumption that $N_e=N_m + N_f$. The inferred male density has a mode at *Nm*=3.01 and limits of an equal-tail 95% credible interval at 1.31 and 5.2. Since the ratio of point estimates for male and female *Nm* values is 20.76, it suggests that male move about 20 times more than females, in agreement with an extreme philopatry of females.

[Figure 2 about here]

DISCUSSION

The results presented here show that ABC can be used to accurately infer parameters under a spatially explicit model using mtDNA sequences or STRs. The use of several summary statistics, each of which contains different information on the expansion process, allows for the simultaneous and accurate estimation of the time since the range expansion, the scaled population size and of the number of migrant genes exchanges between neighbouring demes. Furthermore, the introduction of a Weighted Euclidean Distance giving greater weight to statistics that carry more information on a given parameter is shown to improve the accuracy of the estimation procedure, especially when the number of simulations is limited (Figure 1).

The ABC procedure is robust and estimates parameters consistently well across a broad range of parameter values in the two dimensional stepping stone model. A least-square estimation procedure was proposed earlier to estimate the parameters of a range expansion from mismatch distributions computed from sequence data under an infinite island model (EXCOFFIER 2004). Simulations have shown that this method provides reasonable estimates of τ, but it was found to be highly unreliable for

estimating the migration parameter *M* (data not shown), because the mismatch distribution has a high associated variance for low *M* values and is uninformative for *M* when this parameter is large (*M*>50) (EXCOFFIER 2004). The ability to use other aspects of molecular diversity, both within and between demes considerably improves the estimation of the migration parameter, such that reliable estimates of sex specific dispersal are possible.

As shown above, the accuracy of the estimation increased when more demes were sampled. While sampling from a single deme provided satisfactory estimates for some parameter combinations, it resulted in poor estimates for other combinations, particularly when expansion times were short and migration was low. The improvement due to the use of several demes is linked to the additional information available, but also to the fact that when the total number of demes in the population is large, the gene genealogies of different samples are almost independent after the expansion, thus providing replicate information on the migration process between neighboring demes. Moreover, information on the amount of differences between demes can be incorporated, for instance by using the F_{ST} statistic, while only within deme diversity can be used with a single deme. It should be noted, however, that using information from several demes to estimate a migration parameter assumes that it is a parameter tied to the biology of the species rather than being location specific, and postulates that dispersal patterns will be equivalent at different locations. In that sense the different samples can be considered as replicates of the same process. This assumption is likely to be reasonable when samples are drawn from similar environments. Additionally, the use of several demes in a single estimation procedure assumes that the demes belong to the same expansion wave. Obviously, our results could be erroneous if a species had occupied its present range by a series of

independent range expansions (i.e. from different refuge areas), and if sampled demes were drawn from regions colonized from different sources.

An attractive feature of the ABC method is its ability to handle different types of molecular markers easily, such as DNA sequences or microsatellites, but SNP data could also be used if possible sources of ascertainment bias could be simulated (e.g. WAKELEY *et al*. 2001). One particularly useful application is therefore to compare parameter estimates obtained from different markers, either to assess similarities or any discrepancies that may carry valuable biological information. There is a remarkable concordance in expansion time estimates inferred from mtDNA and STRs for the common vole in Switzerland. These estimates are highly plausible since this suggests the onset of the range expansion occurred after most of the ice cap covering the region during the last glacial maximum had melted by 12 000 years before present and the glaciers had retreated to the highlands (HEWITT 1999). Moreover, from the relatively extensive simulation studies on the performance of our methodology reported in Tables 1 and 2, we see that the coverage property of the posterior distributions is conservative with one million simulations and a tolerance level of 0.1 %, so that we are confident that the estimations obtained from the vole data using the same number of simulation and tolerance level are valid. The existence of glaciation data external to the genetic data is extremely useful to further validate the expansion time estimates. Therefore, the good agreement between these independent data sources suggests that the ABC method may work well when such corroborative data are not available. In contrast to the similarity in estimated expansion times for Northern Swiss vole mtDNA sequences and STRs, migration estimates showed an extreme sex bias in dispersal, with males dispersing at approximately 20 times the rate of females. This extreme difference in the immigration rate between the sexes was

unexpected, although the direction of the bias is completely in agreement with observational data (CLUTTON-BROCK 1989; AARS and IMS 2000; CLOBERT *et al.* 2001; G. HECKEL, unpublished data). Note here that our estimates reflect the effective number of immigrants received by each deme from surrounding unsampled sub-populations, and not the number of genes exchanged between sampled demes, which is in clear contrast with other methods aiming at estimating migration rates (e.g. BEERLI and FELSENSTEIN 2001). However, since the 2DSS model is an approximation of a more continuous distribution of individuals (BARTON and WILSON 1995), the exact delineation of a deme and its neighbourhood size should depend on the pattern of the dispersal distance of individuals, which may not be identical for males and females. In that case, because we expect a quadratic relation ship between the average dispersal distance and the deme size, the number of immigrants (Nm) estimated for males and females may not necessarily apply to the same spatial scale. The 20 fold larger rate of immigration for males compared to female voles could thus be an overestimation if the males were dispersing over larger distances than females. We nevertheless show here how information on nuclear and mtDNA markers can be combined to get estimates of male dispersal, without the need for male-specific markers.

While our study suggests that the ABC procedure is an efficient means of estimating parameters in the 2DSS model, more refined spatial models could be thought of, which could include a number of realistic features such as coastlines or environmental heterogeneity that are likely to be important in shaping the molecular diversity of expanding populations. A simulation model that can include these factors as well as a more progressive range expansion has been recently made available (SPLATCHE, CURRAT *et al.* 2004). Because simulations under this model are considerably slower

than those reported here, its coupling to the ABC procedure is currently difficult, but should be possible with forthcoming increases in computing power. Since the accuracy of the ABC procedure depends in part on the number of simulations used (and hence on simulation time), one interesting challenge for the future will be to determine the benefits of increased model complexity and realism, versus simulation size. Compared to likelihood approaches, the ABC methodology is much easier to implement for complex demographic models, but it does not use as much data, and should thus not be expected to be as accurate. A potential difficulty of ABC approaches is in the choice of summary statistics, and in the definition of the appropriate tolerance level, which are difficult to assess *a priori*. Therefore, simulation studies on the performance of the methodology should be conducted before applying it to observed data, to check for potential biases and for good coverage properties of posterior distributions. Overall, however, this study shows that the ABC procedure should provide a valuable and flexible tool for investigating questions involving range expansions, including those related to sex-biased dispersal or to the history of human settlement.

Acknowledgements

We are grateful to Pierre Berthier and Samuel Neuenschwander for computing assistance, to Sabine Fink and Reto Burri for assistance in the laboratory, and to Arnaud Estoup for helpful comments on the manuscript. This work was supported by a Swiss NSF grant No 3100A0-100800 to LE.

LITERATURE CITED

AARS, J. and R. A. IMS, 2000 Population dynamic and genetic consequences of spatial density-dependent dispersal in patchy populations. American Naturalist **155**: 252-265.

BARTON, N. H. and I. WILSON, 1995 Genealogies and geography. Philos Trans R Soc Lond Bi **349**: 49-59.

BEAUMONT, M. A., 1999 Detecting population expansion and decline using microsatellites. Genetics **153**: 2013-2029.

BEAUMONT, M. A., W. ZHANG and D. BALDING, 2002 Approximate bayesian computation in population genetics. Genetics **162**: 2025-2035.

BEERLI, P., and J. FELSENSTEIN, 2001. Maximum likelihood estimation of a migration matrix and effective population size in *n* subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA, **98**: 4563-4568.

CURRAT, M., N. RAY and L. EXCOFFIER, 2004 SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity. Molecular Ecology Notes **4**: 139-142.

CLOBERT J., E. DANCHIN, A. A. DHONDT and J. D. NICHOLS, 2001 *Dispersal*. Oxford University Press, Oxford.

CLUTTON-BROCK T. H., 1989 Mammalian mating systems. Proceedings of the Royal Society of London, B **236**: 339-372.

ELLEGREN H., 2004 Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics **5**: 435-445.

ESTOUP, A, I. J. WILSON, C. SULLIVAN, J. M. CORNUET and C. MORITZ, 2001. Inferring population history from microsatellite and enzyme data in serially introduced cane toads *Bufo marinus*. Genetics **159**: 1671-1687.

ESTOUP, A., and S. M. CLEGG, 2003 Bayesian inferences on the recent island colonization history by the bird *Zosterops lateralis lateralis*. Molecular Ecology **12**: 657-674.

ESTOUP, A., M. BEAUMONT, F. SENNEDOT, C. MORITZ and J.-M. CORNUET, 2004 Genetic analysis of complex demographic scenarios : spatially expanding populations of the cane toad, *Bufo marinus*. Evolution (in press).

EXCOFFIER, L., 2004 Patterns of DNA sequence diversity and genetic structure after a range expansion: Lessons from the infinite-island model. Molecular Ecology 13: 853-864.

FINK S., L. EXCOFFIER and G. HECKEL, 2004 Mitochondrial gene diversity in the common vole *Microtus arvalis* shaped by historical divergence and local adaptations. Molecular Ecology 13: 3501-3514.

FU, Y. X. and W. H. LI, 1997 Estimating the age of the common ancestor of a sample of DNA sequences. Mol Biol Evol **14:**195-199.

GOLDSTEIN D. B., G. W ROEMER, D. A. SMITH, D. E. REICH, A. BERGMAN, and R. K. WAYNE, 1999 The use of microsatellite variation to infer population structure and demographic history in a natural model system. Genetics **151**: 797-801.

HAUSSER J., 1995 Säugetiere der Schweiz: Verbreitung, Biologie, Oekologie. Birkhäuser Verlag, Basel.

HAYNES S, M. JAAROLA, and J. B. SEARLE, 2003 Phylogeography of the common vole (*Microtus arvalis*) with particular emphasis on the colonization of the Orkney archipelago. Molecular Ecology **12**: 951-956.

HECKEL G., R. BURRI, S. FINK, J.-F. DESMET, and L. EXCOFFIER. Southwestward colonization of Europe by the common vole *Microtus arvalis*, Evolution, submitted

HEWITT G. M., 1999 Post-glacial re-colonization of European biota. Biological Journal of the Linnean Society **68**: 87-112.

HEWITT G. M., 2000 The genetic legacy of the Quaternary ice ages. Nature **405:** 907-913.

HUDSON, R. R. 1990 Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. oxford: Oxford University Press

JARNE P., and P .J. L LAGODA, 1996 Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution **11**: 424-429.

MITCHELL-JONES A. J, G. AMORI, W. BOGDANOWICZ, B. KRYSTUFEK, P. J. H. REINJNDERS, F. SPITZENBERGER, M. STUBBE, J. B. M. THISSEN, V. VOHRALIK & J. ZIMA 1999 *The atlas of European mammals*, T. & A.D. Poyser, London.

NIELSEN R. and J. WAKELY, 2001 Distinguishing migration from isolation: A Markov chain Monte-Carlo approach. Genetics **158**: 885-896.

PERRIN N., and V. MAZALOV, 2000 Local competition, inbreeding, and the evolution of sex-biased dispersal. American Naturalist **155**: 116-127

PRITCHARD, J. K, M. T. SEIELSTAD, A. PEREZ-LEZAUN and M. W. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol **16:** 1791-1798.

RAY, N., M. CURRAT, and L. EXCOFFIER, 2003 Intra-deme molecular diversity in spatially expanding populations. Mol Biol Evol **20:** 76-86.

TAVARÉ, S., D. J. BALDING, R. C. GRIFFITHS and P. DONNELY, 1997 Inferring coalescence times from DNA sequence data. Genetics **145**: 505-518.

WAKELEY, J., R. NIELSEN, S. N. LIU-CORDERO, K. ARDLIE, 2001 The discovery of single-nucleotide polymorphisms-- and inferences about human demographic history. Am J Hum Genet **69**: 1332-47

WANG J., and M. C. WHITLOCK, 2003 Estimating effective population size and migration rates from genetic samples over space and time. Genetics **163**: 429-446.

Table 1:Range expansion parameters estimated from a single deme for samples of mtDNA sequences and 10 STR loci.

τ	$\bar{\hat{\tau}}$	$\hat{\tau}$ 2.5% quantile	$\hat{\tau}$ 97.5% quantile	95% coverage	θ	$\bar{\hat{\theta}}$	$\hat{\theta}$ 2.5% quantile	$\hat{\theta}$ 97.5% quantile	95% coverage	M	$\bar{\hat{M}}$	$\hat{M}$ 2.5% quantile	$\hat{M}$ 97.5% quantile	95% coverage
a) DNA sequence (300 bp)														
1	10.9	7.3	14.1	1000	**0.1**	0.3	0.1	0.7	1000	**1**	0.1	0.1	0.4	1000
1	8.2	3.8	12.5	1000	**0.1**	0.6	0.2	1.3	1000	**10**	0.3	0.1	1.2	1000
1	1.8	0.8	4.7	1000	**0.1**	0.3	0.1	0.9	1000	**100**	62.5	0.6	263.6	1000
1	9.0	3.2	14.1	1000	**1**	0.8	0.2	1.9	1000	**1**	0.3	0.1	0.9	1000
1	6.3	2.0	11.6	999	**1**	1.0	0.2	2.7	996	**10**	1.4	0.1	11.5	1000
1	2.4	0.9	6.5	999	**1**	0.6	0.1	2.9	999	**100**	71.1	0.3	302.5	1000
10	13.0	8.4	18.5	967	**0.1**	0.1	0.0	0.3	1000	**1**	0.6	0.2	1.5	980
10	14.0	8.9	22.3	922	**0.1**	0.3	0.1	1.3	1000	**10**	4.6	0.7	12.8	1000
10	9.8	7.6	11.9	986	**0.1**	0.4	0.2	0.9	994	**100**	78.3	31.4	138.7	993
10	12.3	7.0	18.7	975	**1**	0.6	0.1	2.0	1000	**1**	1.0	0.1	3.2	943
10	14.1	9.2	21.7	941	**1**	0.7	0.1	2.9	1000	**10**	8.4	0.8	26.2	997
10	10.3	8.6	12.1	987	**1**	0.6	0.4	0.9	1000	**100**	124.8	66.1	218.4	986

τ	$\bar{\hat{\tau}}$	$\hat{\tau}$ 2.5% quantile	$\hat{\tau}$ 97.5% quantile	95% coverage	θ	$\bar{\hat{\theta}}$	$\hat{\theta}$ 2.5% quantile	$\hat{\theta}$ 97.5% quantile	95% coverage	M	$\bar{\hat{M}}$	$\hat{M}$ 2.5% quantile	$\hat{M}$ 97.5% quantile	95% coverage
b) 10 STR loci														
20	34.7	21.3	49.8	983	**5**	1.7	0.5	3.5	1000	**1**	0.9	0.3	2.2	1000
20	34.3	22.3	48.8	999	**5**	2.1	0.9	3.9	1000	**10**	2.4	1.4	4.4	1000
20	19.4	15.6	24.8	1000	**5**	3.7	2.3	5.7	1000	**100**	65.3	7.1	213.7	1000
20	32.2	16.6	49.4	980	**10**	1.5	0.5	2.8	1000	**1**	1.9	0.7	4.6	1000
20	31.4	18.6	46.2	1000	**10**	2.9	1.4	4.9	1000	**10**	3.6	1.7	9.6	1000
20	21.1	17.5	27.3	1000	**10**	4.4	2.8	6.6	1000	**100**	84.9	6.9	254.7	1000
40	47.4	28.6	75.4	989	**5**	1.3	0.6	2.4	1000	**1**	1.8	0.6	3.1	1000
40	55.9	39.1	75.9	982	**5**	3.1	2.0	4.3	1000	**10**	3.9	1.3	8.6	1000
40	36.6	29.8	46.3	1000	**5**	9.5	7.1	12.4	1000	**100**	30.2	3.4	118.1	1000
40	46.6	27.5	72.4	992	**10**	1.7	0.9	3.0	1000	**1**	2.2	1.0	4.4	1000
40	53.8	36.7	74.2	1000	**10**	4.4	2.8	7.3	1000	**10**	3.7	1.5	8.7	1000
40	38.4	30.5	49.1	1000	**10**	9.5	7.1	12.6	1000	**100**	52.3	4.3	186.8	1000

Average estimation, 2.5 and 97.5 quantile values of the distributions of estimates from 1000 simulated samples.

95% coverage reports the number of simulated samples (among 1000) for which the true parameter lies within the 95% credible interval estimated from the posterior distribution.

Table 2: Range expansion parameters estimated from three demes for samples of mtDNA sequences and 10 STR loci.

τ	$\bar{\hat{\tau}}$	$\hat{\tau}$ 2.5% quantile	$\hat{\tau}$ 97.5% quantile	95% coverage	θ	$\bar{\hat{\theta}}$	$\hat{\theta}$ 2.5% quantile	$\hat{\theta}$ 97.5% quantile	95% coverage	M	$\bar{\hat{M}}$	$\hat{M}$ 2.5% quantile	$\hat{M}$ 97.5% quantile	95% coverage
a) DNA sequence (300 bp)														
1	1.9	0.2	6.0	978	**0.1**	0.1	0.0	0.3	1000	**1**	0.7	0.1	2.9	986
1	1.0	0.4	1.7	992	**0.1**	0.3	0.1	0.6	1000	**10**	5.3	0.9	17.4	994
1	1.1	0.7	1.5	998	**0.1**	0.2	0.0	0.5	1000	**100**	123.8	32.8	235.8	1000
1	1.4	0.3	4.1	990	**1**	0.5	0.1	1.2	986	**1**	2.1	0.3	9.5	979
1	1.0	0.5	1.8	976	**1**	0.6	0.1	1.7	991	**10**	12.3	1.1	48.1	998
1	1.1	0.7	1.7	993	**1**	0.4	0.1	2.0	973	**100**	107.1	9.2	258.8	997
10	11.2	7.5	15.5	958	**0.1**	0.1	0.0	0.3	1000	**1**	0.9	0.2	2.0	980
10	10.5	8.3	13.3	965	**0.1**	0.1	0.0	0.3	1000	**10**	9.9	4.6	15.5	989
10	9.9	8.3	11.6	988	**0.1**	0.3	0.1	0.8	963	**100**	92.2	51.2	150.3	995
10	11.4	6.9	16.4	978	**1**	0.6	0.1	1.8	967	**1**	1.1	0.2	2.8	989
10	10.3	8.3	13.0	988	**1**	0.4	0.1	1.1	953	**10**	13.0	6.0	21.0	999
10	10.0	8.8	11.4	997	**1**	1.0	0.3	3.0	995	**100**	103.9	59.9	153.8	997

τ	$\bar{\hat{\tau}}$	$\hat{\tau}$ 2.5% quantile	$\hat{\tau}$ 97.5% quantile	95% coverage	θ	$\bar{\hat{\theta}}$	$\hat{\theta}$ 2.5% quantile	$\hat{\theta}$ 97.5% quantile	95% coverage	M	$\bar{\hat{M}}$	$\hat{M}$ 2.5% quantile	$\hat{M}$ 97.5% quantile	95% coverage
b) 10 STR loci														
20	21.8	13.7	32.9	974	**5**	3.9	1.4	7.0	990	**1**	0.7	0.1	2.1	948
20	19.5	14.1	27.0	1000	**5**	3.9	0.6	9.8	1000	**10**	6.8	1.1	22.1	981
20	17.7	14.5	21.1	999	**5**	4.2	0.9	15.0	999	**100**	125.7	25.4	252.1	980
20	21.5	13.7	32.3	978	**10**	7.2	3.1	12.4	997	**1**	0.5	0.1	1.6	932
20	18.8	13.0	26.8	999	**10**	9.0	1.5	21.1	1000	**10**	4.9	0.7	18.8	936
20	19.7	16.1	24.1	1000	**10**	5.2	1.2	17.0	1000	**100**	143.1	31.5	279.2	976
40	44.7	27.5	67.5	960	**5**	3.7	1.3	7.9	981	**1**	0.9	0.3	2.4	961
40	41.9	31.1	54.0	1000	**5**	3.5	0.7	8.1	999	**10**	9.5	2.7	23.4	981
40	32.8	26.2	39.2	998	**5**	7.3	1.1	26.4	999	**100**	97.3	27.1	199.1	986
40	44.1	27.9	66.2	967	**10**	6.7	2.4	14.7	987	**1**	0.7	0.2	1.9	962
40	39.5	28.5	53.6	1000	**10**	7.5	1.7	17.0	1000	**10**	7.5	1.4	24.0	986
40	35.3	27.9	42.8	998	**10**	8.9	1.5	33.1	1000	**100**	124.9	36.7	239.3	991

See table 1 for details.

Table 3: Range expansion parameter estimates and the number of immigrant genes (*Nm*) per deme per generation with 95% credible intervals (C.I.) for five vole populations in northern Switzerland, for samples of mtDNA sequences and STR loci.

	$\hat{\tau}$	95% C.I.	$\hat{\theta}$ [1]	95% C.I.	M [2]	95% C.I.	Nm	95% C.I.
mtDNA (321 bp)	4.8	2.8-8.0	1.8	0.1-6.1	0.29	0.02-2.5	0.15	0.01-1.25
STR (12 loci)	23	16.4-34	0.08	0.02-0.45	13.2	5.6-20.4	3.3	1.4-5.1

[1] θ estimates are presented for the whole chromosome rather than per site (bp or STR locus).

[2] M is equal to two times the number of immigrant females for mtDNA data, while it is equal to four times the total number (males and females) of immigrants for nuclear STR data.

Figure 1: A plot of RMSE for estimates of the migration parameter *M* versus tolerance (Pδ). Tolerance represents here the fraction of simulations retained for parameter estimation. Estimates using the standard ABC method are shown as white squares, and those using the modified Weighted Euclidean Distance are shown as black diamonds. One million simulations were used for each estimation. Bootstrapped standard errors were calculated on 10 000 replicates, and are reported here around point estimates.

Figure 2: Plots of the posterior densities of the number of immigrant genes (*Nm*) per deme per generation obtained from the analysis of 5 North-Western Swiss vole populations. The posterior density estimated using STRs is shown as a thin line, the posterior density estimated using mtDNA is shown as a thick line, and the estimated posterior density of the males obtained as a convolution of mtDNA and STR densities is shown as a dashed line.

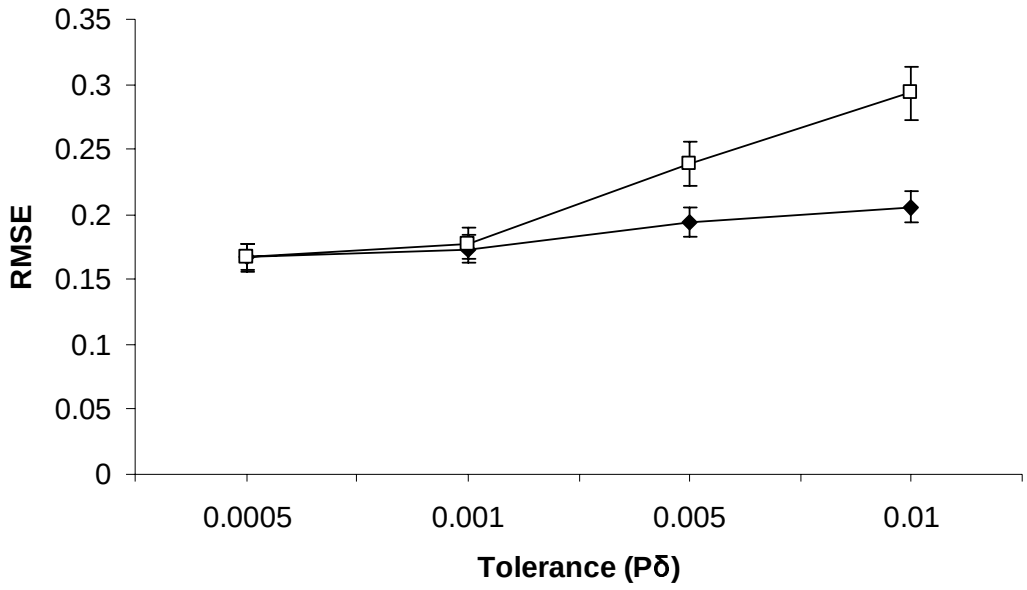
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
RMSE
0.0005
0.001
0.005
0.01
Tolerance (Pδ)

Figure 1

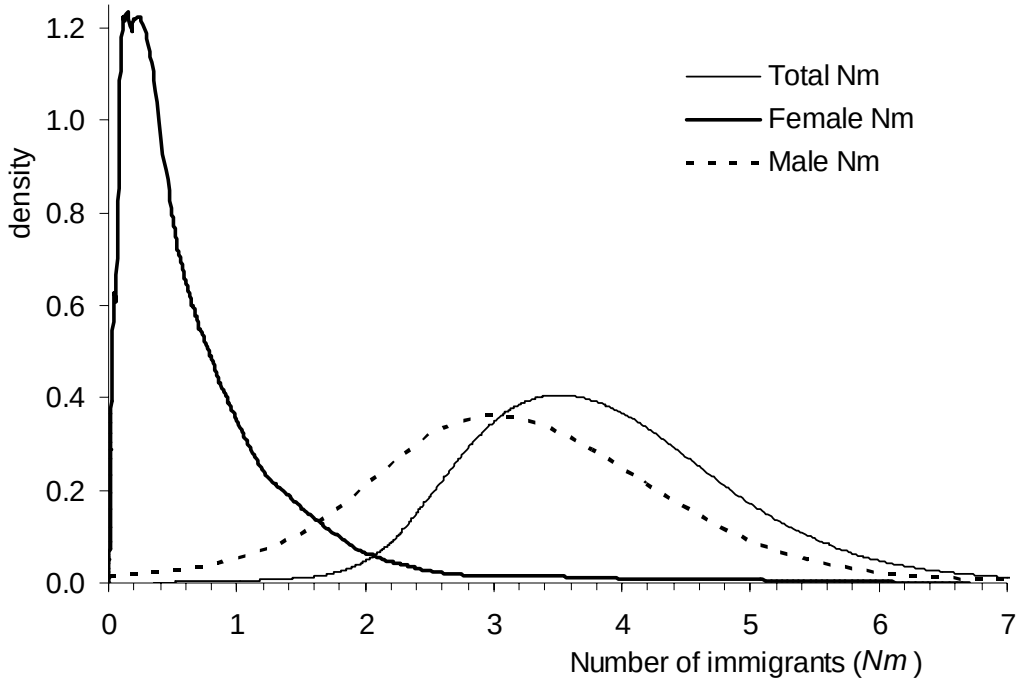
1.2
1.0
0.8
0.6
0.4
0.2
0.0
density
0
1
2
3
4
5
6
7
Number of immigrants (Nm)
Total Nm
Female Nm
Male Nm

Figure 2