Chapter 14

Joint Determination of Topology, Divergence Time and Immigration in Population Trees[1]

Mark A. Beaumont

It is often hoped that population genetics will be a strong partner in archaeological and linguistic analysis of human prehistory. There are, however, many challenges, not least because there is rather little information in most data with which to draw strong conclusions. Computer-intensive statistical methods have been developed to extract as much information from the data as possible, and to provide a flexible framework within which complex models of population history can be handled. To be able to advance further, we need to analyse autosomal DNA variation at many loci. Current computer-intensive methods that aim to analyse the data fully and accurately begin to fail when faced with many loci and complex models. There has been a move towards the development of more approximate computer-intensive methods. This paper describes a particular class of models that has wide applicability, where populations diverge genetically though time, influenced by random genetic drift and migration. I describe an approximate Bayesian method that uses summary statistics measured from microsatellite loci to make inferences about demographic parameters in two- and three-population models. The method can be used to infer effective sizes of current and ancestral populations, immigration rates, splitting times and tree topology (in the three-population case). A novel method, based on categorical regression, for model selection is introduced, and used to infer tree topologies. Comparisons are made with a full-likelihood method for two populations developed by Hey and Nielsen, and it appears that the method gives comparable results. Multiple simulated test data sets are analysed with the three-population model, and it is concluded that in the presence of immigration the ability to make strong inferences about the population tree topology is quite weak. The categorical regression method for model selection is demonstrated to be substantially more efficient than simple rejection. I analyse data sets of 19 microsatellite loci from Channel Island foxes, and 329 microsatellite loci from three human populations. In the case of the foxes no strong conclusions about population tree topology can be made. With the human data, much stronger inferences about topology can be made. Overall, however, there appears to be little scope for accurate inference of demographic parameters with microsatellite data in the face of immigration, even when large numbers of microsatellite loci are used.

1. Introduction

Population genetic analysis has long held out the prospect that it can greatly enhance and complement the insights into human prehistory gleaned from archaeological and linguistic studies. Although there have been some successes, particularly with respect to the support given for the idea that modern humans

emerged from Africa relatively recently, with hindsight we can see that current studies represent the steady trudge up the foothills of a mountain whose grand vistas are yet to open up. There are a number of reasons for this. Most pertinent is that there is actually rather little information, at least of the kind relevant to archaeologists, in much of the genetic material typically analysed. Secondly, those data that are of potential value, such as autosomal SNPs and multilocus autosomal sequences, are currently very hard to analyse effectively.

Although population genetics has a rich history of statistical analysis, most of the methods originally developed were of limited application, and not designed for the revolution in molecular genetics underway throughout the 1980s. The conceptual framework for analysing molecular genetic data as efficiently as possible was laid down in the 1990s, when computer-intensive statistical methods, originating in physics, were first applied to population genetic problems (Griffiths & Tavaré 1994; Kuhner *et al.* 1995; Wilson & Balding 1998). It became clearer how to probabilistically model genetic data, taking both Bayesian and likelihood-based frequentist approaches. These methods have an inherent flexibility, not readily accessible in the moment-based statistical techniques that had traditionally been used, and potentially allow for highly complex demographic and evolutionary scenarios to be modelled.

One goal of these recent techniques is to try to better understand the demographic prehistory of human populations by inferring the parameters in specific models. One particular class of models has been motivated by the idea that, if different geographic regions have been populated relatively recently from a common ancestral stock, the genetic differences between populations may tell us something about the time since the regions were colonized, the rate of migration between the populations, their sizes, and changes in population size associated with colonization. More concretely, when, for example, one group of humans from a certain population colonizes another region it is as if the original population has split into two. Initially the gene frequencies may be similar, depending on the size of the founding group, but then, due to random changes in gene frequency from generation to generation (random genetic drift) the gene frequencies in the populations will steadily diverge through time. Migration between the two populations may keep the gene frequencies similar. After a while, depending on the rate of migration and population sizes, the gene frequencies become independent of the time of the colonization event, and the populations are said to be in immigration-drift balance. In this case we can no longer say anything about the time of population divergence or ancestral population size, but can still say something about migration rates and population sizes. These processes can be encapsulated in a population genetic model, and we can then, given some data, try to infer the parameters of the model.

The genetic analysis of geographically structured populations has tended to be influenced by the two underlying conceptual models alluded to above. In one scenario the observed pattern of genetic variation arises from a process of immigration, mutation, and random genetic drift at equilibrium (Wright 1931; 1937). In the other, the pattern is generated by transient historical phenomena such as the splitting of populations and their subsequent genetic divergence. For this latter case, typically, moment-based methods have been used to obtain estimates of the time of population divergence. The topology of the population tree has often been obtained from ordination methods, such as neighbour-joining (Saitou & Nei 1987).

Since the middle of the 1990s microsatellites have often been used to infer population topologies, and moment-based methods that incorporated appropriate mutational models were developed to infer times of population divergence (e.g. Goldstein *et al.* 1995). A comparison of these methods on simulated data sets was carried out by Takezaki & Nei (1996), and they concluded that many hundreds of loci would be needed to infer parameters with any certainty. They also noted that drift-based estimators were better at recovering the topology of the population tree, while estimators based on a mutation model were better at obtaining the times of population divergence.

Ideally, likelihood-based procedures should be more efficient in extracting the information in the data available to infer topologies and branch lengths. Models in which populations diverge through time have been the subject of early attempts at likelihood-based analysis to infer demographic parameters (Cavalli-Sforza & Edwards 1967). A Bayesian method has been developed for single or linked microsatellite loci, and single locus sequence data modelled through the infinite-sites approximation (the program BATWING: Wilson *et al.* 2003). In BATWING, Bayesian computation is performed via Markov chain Monte Carlo (MCMC) simulation, and, in principle, the procedure can be used for any number of populations, although in practice convergence of the MCMC restricts the range of application of this approach.

A useful advance has been the introduction of probabilistic models that combine both equilibrium and non-equilibrium features (Nielsen & Wakeley 2001; Hey & Nielsen 2004). As with the method of Wilson *et al.* (2003), inference under this model is

performed via MCMC to obtain relative likelihoods for different parameter values. This approach can infer immigration rates, times of population splitting, and population sizes for a pair of populations. The recent improvement described in Hey & Nielsen (2004), implemented in the IM ('Isolation with Migration') program, allows the analysis of multi-locus data evolving under a number of different mutational models (sequences, SNPs, microsatellites).

A natural extension of the method is to consider larger numbers of populations that can diverge in a bifurcating way (as in Wilson *et al.* 2003), in the presence of gene flow. However, while very powerful, it has to be recognized that, as with many MCMC-based genealogical methods, convergence can be quite slow, particularly for large data sets, and it may be difficult to extend the computational approach in Hey & Nielsen (2004) to consider larger numbers of populations. There has been a general recognition (Li & Stephens 2003) that methods of genealogical analysis based on MCMC or importance sampling (IS) are typically restricted to small data sets. The reason for this is that the likelihoods can only be written down for a single genealogical history, but many histories are compatible with any given data set. Thus the dimensionality of the problem is already very large for even a small amount of data, and MCMC or importance sampling methods struggle to sample from the even vaster space associated with larger data sets.

In recognition of this, various other approximate methods have been suggested that allow the flexibility of Bayesian and likelihood-based inference, while at the same time allowing for larger data sets to be considered (Hudson 2001; McVean *et al.* 2002; Li & Stephens 2003).

Given the weak relationship between the patterns found in population genetic data, and the evolutionary and demographic history we typically wish to uncover, the future of these methods surely lies in analysis of multiple nuclear sequences. Such data sets are becoming more widely available in a wide variety of organisms (Jennings & Edwards 2005). However, it is evident from these data that the effect of recombination is ubiquitous, and this vitiates the assumptions inherent in the use of the IM program, and may necessitate discarding significant amounts of sequence data (Hey & Nielsen 2004). Ideally, recombination needs to be included in the model, but this would greatly increase the dimensionality of the problem, and, realistically, can probably only be practicably addressed using approximate likelihood-based methods.

It seems apparent that, after the first flush of promise of MCMC and IS approaches in the late 1990s, the future of population genetic data analysis lies in the exploration and development of alternative, approximate, statistical methods that also have the flexibility found in Bayesian and likelihood-based analysis. The motivation of this paper is to explore how easy it is to tackle similar problems to those considered by Hey & Nielsen (2004) using an approximate method based on summary statistics. This approach, first proposed in population genetics context by Pritchard *et al.* (1999), has come to be called 'Approximate Bayesian Computation' (ABC) (Beaumont *et al*. 2002; Marjoram *et al.* 2003). The aim of this study is to compare the method with that of Hey & Nielsen (2004) on a pair of populations, and then to examine the behaviour of an extension of the method to three populations. In this latter case the tree topology becomes a parameter that also needs to be inferred.

2. Models and methods

The ABC approach to statistical inference does not require an explicit likelihood function. All that it requires is a probabilistic simulation program that can generate data sets with the same attributes as those found in the 'real' data set, obtained from nature. The goal is to capture these attributes via summary statistics and thereby make inferences about parameters in the simulation model. A good description of the method is in Excoffier *et al.* (2005). Below, I first describe the demographic and genetic models whose parameters we wish to infer; I then describe the ABC method; and finally describe the summary statistics that are calculated from the real and simulated data.

2.1. The demographic and genetic model

The models described here are of two or three populations. Data were simulated using a coalescent simulation, modified from the method described in Beaumont & Nichols (1996). A single-step microsatellite mutation model was used for these simulations, although the method allows for more general mutation models to be considered (Pritchard *et al.* 1999).

In the case of two populations the model was identical to that in, for example, Nielsen & Wakeley (2001) and Hey & Nielsen (2004). The treatment in these papers is in terms of scaled parameters. Here I describe the model in terms of the natural parameters (Fig. 14.1), but in comparisons between the ABC method and the IM program will also give the scaled parameters below. This is a model of an ancestral population of size N_A that splits at some time T in the past to give two populations of size N_1 and N_2. Since the time of splitting (looking forward in time) there has been immigration at rates m_1 and m_2, where immigrants into one population are drawn from the other population.

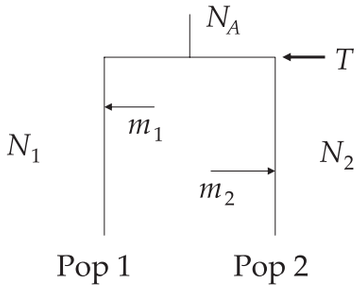

Figure 14.1. *Illustration of the two-population model. See text for description of parameters.*

Variation in mutation rate among loci was modelled in a hierarchical way following Storz & Beaumont (2002). The mutation rates at each locus are assumed to be drawn from a lognormal distribution with (on the $\log_{10}$ scale) mean of μ and standard deviation σ. Inference were made on these parameters rather than the mutation rates at individual loci. Note, this is somewhat different from the method of modelling variation in mutation rates among loci adopted by Hey & Nielsen (2004). The scaled parameters in Hey & Nielsen (2004) are $t = \mu T$, $\theta_1 = 2N_1\mu$, $\theta_2 = 2N_2\mu$, $\theta_A = 2N_A\mu$, $\mathcal{M}_1 = m_1/\mu$, $\mathcal{M}_2 = m_2/\mu$.

In the case of three populations we have parameters N_1, N_2, and N_3 for the sizes of the current populations, N_{A1} for the size of the most recent ancestral population, and N_{A2} for the common ancestral population size (Fig. 14.2). The time of the most recent split is T_1 and the time of the most ancient split is T_2. The immigration rates for the current populations are m_1, m_2, and m_3. The immigration rate for the most recent ancestral population is m_a. When there are three populations (i.e. in the interval between the most recent split and the current time) the immigrants into one population are drawn with equal probability from the other two populations (thus an island model is assumed). Variation in mutation rate among loci was modelled as for the two-population case. In addition for three populations there are 3 possible tree topologies, given indicators (1, 2, 3) for ordered populations ((POP1, POP2), POP3), ((POP1, POP3), POP2), ((POP2, POP3), POP1). Thus, overall, for the two-population case there are 8 parameters that can be inferred, and for the three-population case there are 14 parameters. Further parameters could be included in the three-population case. For example, immigration rate and effective size could change at the time of the most recent split for the population not involved in the split (i.e. N_{A3} and M_3 could change at the time of the split: see Fig. 14.2), which might be reasonable if this vicariance event also had an effect on the other population.

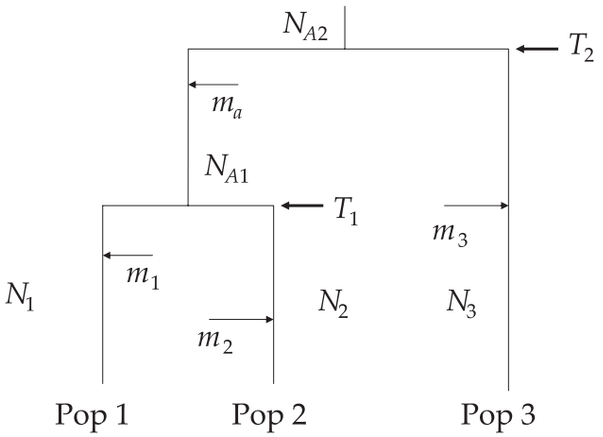

Figure 14.2. *Illustration of the three-population model. See text for description of parameters.*

Also, a migration matrix could be used during the period of three populations, instead of assuming an island model. These enhancements would then lead to a 19 parameter model.

2.2. ABC approach to genealogical inference

In the simplest Bayesian calculation we have a probabilistic model that gives the joint distribution of an (unobserved) parameter Φ and some measurement that could be obtained from the data, S. Typically we make this calculation via the likelihood $P(S|\Phi)$, which gives the probability distribution of any value of S, given a value of Φ, and prior distribution for the parameter, $P(\Phi)$. The product of these gives the joint distribution $P(S, \Phi)$. If we have a specific observation, s, we then want to know the distribution of parameter values given this observation, $P(\Phi|S = s)$. This is the posterior distribution and can, in principle be computed as $P(\Phi, S)/P(S)$. The ABC approach simply involves simulating parameters from the joint distribution, $P(\Phi, S)$, and then using some density estimation method to obtain an approximation to this conditional distribution (Fig. 14.3). Simulation from the joint distribution can be done in two steps: first simulate a parameter values from the prior distribution, and then, using this value, simulate from the likelihood; repeat as many times as needed. One aspect, which makes the method attractive and flexible, is that there is no need to have (or manipulate in any way) explicit analytical expressions for $P(\Phi, S)$, $P(\Phi)$, or $P(S|\Phi)$. One simply needs some method of simulating from these distributions. Typically the data will be summarized not by one value, but by a number of summary statistics. The general idea is to use as many summary statistics as possible, so that it is possible to replace the data with these summaries

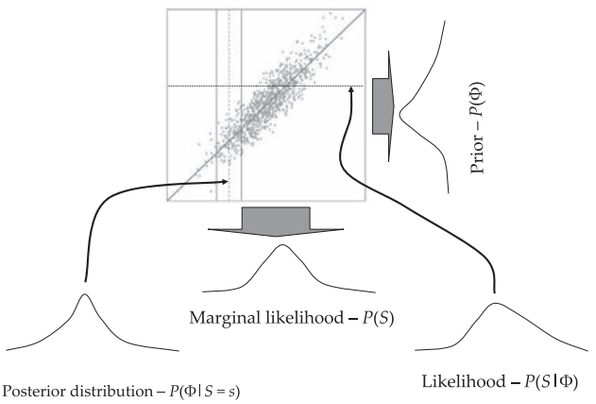

Figure 14.3. *This figure illustrates the basic ABC approach. The points in the figure are simulated from the joint distribution of parameter values and data. The aim of the ABC approach is to take a 'slice' through the distribution at a point corresponding to the measurement obtained from the real data, and thereby compute the conditional distribution of parameter values.*

and still obtain similar posterior densities to those that would be obtained with a method that could use all the data (i.e. the summary statistics are 'sufficient' in the statistical sense).

In the original specification of Pritchard *et al.* (1999), the conditional distributions were calculated by simply taking all the points simulated from the joint distribution that had summary statistics within some narrow interval around those observed in the data (as illustrated in Fig. 14.3). As pointed out in Beaumont *et al.* (2002), there are other potentially more efficient methods of conditional density estimation, and they proposed a method based on local linear regression. The main difference between the two methods is that, with the earlier method, if the interval for accepting the points (the region within the two lines on either side of the dotted line in Fig. 14.3) is made large enough the prior will then be recovered (since the parameter values are drawn from the prior). This is not the case for the regression method, and if the joint distribution is multivariate normal, as illustrated in the figure, then a good estimate of the true posterior distribution will be recovered by accepting all the points. The analysis presented in this paper is based on the regression method, which is summarized in Appendix 14.1.

2.3. Summary statistics

A weakness of ABC methods is that there is currently no objective 'rule' for choosing summary statistics, and therefore the eventual choice is somewhat arbitrary. Another problem to consider is the explosion of summary statistics that can potentially occur as one considers an increasing number of populations. The summary statistics described here were chosen primarily because of their history of use in previous (moment-based) studies of microsatellite data (e.g. Takezaki & Nei 1996; King *et al.* 2000):

1. heterozygosity in each population $n/(n-1)(1-\sum x_i^2)$;
2. sample variance of allele length in each population;
3. number of alleles in each population;
4. heterozygosity for each pair of populations pooled together;
5. variance for each pair of populations pooled together;
6. number of alleles for each pair of populations pooled together.

These were measured for each locus and then averaged. If, for any locus, the sample size in a population was ≤1 for any population, the summary statistics for that locus were not included in the average for the population, or pair of populations. In total, for two-population models there were 9 summary statistics, and for three-population models there were 18.

2.4. Model selection

Making inferences about the posterior probability of particular models, poses some challenges for computational methods such as MCMC, although the use of reversible-jump MCMC has proved very useful (Green 1995). As pointed out in Pritchard *et al.* (1999), it is relatively straightforward in principle to estimate the posterior probability of particular models using approximate methods based on summary statistics: we estimate the marginal probability of the summary statistics under model M_1, $\pi_{M1}(S = s)$, by simply counting up the proportion of simulated points that are within our tolerance region of the target summary statistics $\| S_i - s \| < \delta$. Two models can then be compared as

$$\frac{\hat{\pi}_{M1}(S = s)}{\hat{\pi}_{M2}(S = s)}.$$

However, as with parameter estimation, the use of straightforward rejection is potentially inefficient, and may be improved with kernel-based methods for estimation of the density.

However, another alternative, explored for the first time in this article, is to directly estimate the posterior probability of a model itself rather than to do so indirectly via comparison of estimates of $\pi_{Mi}(S = s)$. This can be straightforwardly achieved in the regression framework by treating the model indicator as a categorical variable Y that can take values from

$(1, \cdots, n_M)$. We can then estimate the coefficients β in a multinomial logit model in which

$$P(Y = j \mid S) = \frac{\exp(\beta_j^T S)}{\sum_{i=1}^{M} \exp(\beta_i^T S)}$$

and thereby obtain an estimate of $P(Y = j \mid S = s)$. This can be performed using weighted regression, as described above. The method used in this paper is implemented in the VGAM package by Thomas Yee under R (http://www.stat.auckland.ac.nz/~yee).

2.5. Comparison with IM

The IM package was used to compare posterior distributions from a full-data method with those obtained under the ABC method for 5 test simulations. Simulated data sets consisted of samples of size 50 chromosomes in each population scored for 10 microsatellite loci. The parameters used to simulate the data sets were $t = 4$, $\theta_1 = 0.5$, $\theta_2 = 2$, $\theta_A = 10$, $\mathcal{M}_1 = 4$, $\mathcal{M}_2 = 1$.

For the simulations described here the IM version of (11/12/04) was used. Priors for θ_1, θ_2 and θ_A were chosen to be uniform on the range (0,30), for $\mathcal{M}_1$ and $\mathcal{M}_2$ uniform (0,10), and for t uniform (0,8). Two independent simulations were run, each for 5×10^7 updates after an initial discarded burn-in of 500,000 updates. The Metropolis-coupling option was not used.

In order to make comparisons between methods, for the ABC simulations μ was given a point prior of 5×10^{-4}, and the priors for the other parameters chosen so that the priors on the scaled parameters were identical to those of IM. Variability in mutation rate could not be matched identically, because of the different method used to represent this. In the ABC analysis the prior for σ was a normal distribution (as in Storz & Beaumont 2002), truncated at 0, with mean 0 and standard deviation 0.1 (compatible with up to a 100-fold variation in mutation rates among loci).

The ABC analysis of the data was similar to that described in Beaumont *et al.* (2002) except that, since the parameters are distributed along uniform bounds, a logistic transformation was used prior to the regression adjustment, and then the regression-adjusted parameter values were back-transformed. Five hundred thousand points from the joint distribution of parameters and summary statistics were simulated, and the 2000 points closest to each target set of summary statistics were used for regression-adjustment ($P_\delta = 0.004$) (see Appendix 14.1).

2.6. Simulations for three populations

Two sets of test data was generated, consisting of 100 independent simulations of respectively 50 and 500 loci with a sample size of 50 chromosomes per population. For ease of exposition, because scaling of parameters can be confusing in complex demographic models, in the description below I fix the mutation rate at a notional $\mu = 5 \times 10^{-4}$ both in the generation of test data sets and in the analysis. To generate the test data the following parameters were used: $\sigma = 0$ (i.e. no variation in mutation rate among loci), $N_1 = 1000$, $N_2 = 1000$, $N_3 = 1000$, $N_{A1} = 1000$, $N_{A2} = 1000$, $T_1 = 100$, $T_2 = 1100$, $m_1 = m_2 = m_3 = m_a = 0.0005$.

For the ABC analysis, three different sets of prior distributions were used to analyse the test data with 50 loci. Each of these was replicated twice, leading to six independent sets of 500,000 points. Of these, the 10,000 points with the shortest Euclidean distance to each target set of summary statistics were used in the regression adjustment ($P_\delta = 0.02$). In the first set of replicates a point prior equal to that in the simulated data was set for the divergence times; in the second set, the immigration rates were fixed at the 'true' values; and in the third set the demographic parameters were free to vary. When not fixed, the following priors were used: a uniform prior on (0, 10,000) for N_1, N_2, N_3, N_{A1}, and N_{A2}, a uniform prior on (0, 10,000) for T_1 and $T_2 - T_1$, a uniform prior on (0, 0.005) for the mutation rates, and a uniform prior on the three different topologies. The mean mutation rate was fixed at $\mu = 5 \times 10^{-4}$, with standard deviation among loci at $\sigma = 0.1$, as for the two-population case. For the simulation tests with 500 loci, the same settings were used to simulate and analyse the data as for 50 loci, but in this case only the scenario without point priors for immigration or divergence time was studied.

3. Results

3.1. Two populations: a comparison of ABC and MCMC

Visual inspection of the output from the IM analysis suggested that reasonable convergence had been achieved, although mixing for t was relatively poor, and there was some variation in the two independent chains. It is possible that improved performance could have been obtained by choosing the Metropolis-coupling option, but earlier tests using Metropolis-coupling with 5 parallel chains were not encouraging. This is not to conclude that Metropolis-coupling is not useful, but it would appear that some effort is needed in optimizing the heating-parameters and numbers of chains. Each IM simulation took around 45 hours to run on a Xeon 3.2Ghz (Nocona) computer (maximally optimized using an Intel compiler under Linux). In the ABC analysis the 500,000 points were simulated in 20 minutes, and the same set was then used in the analysis of all simulated data sets.

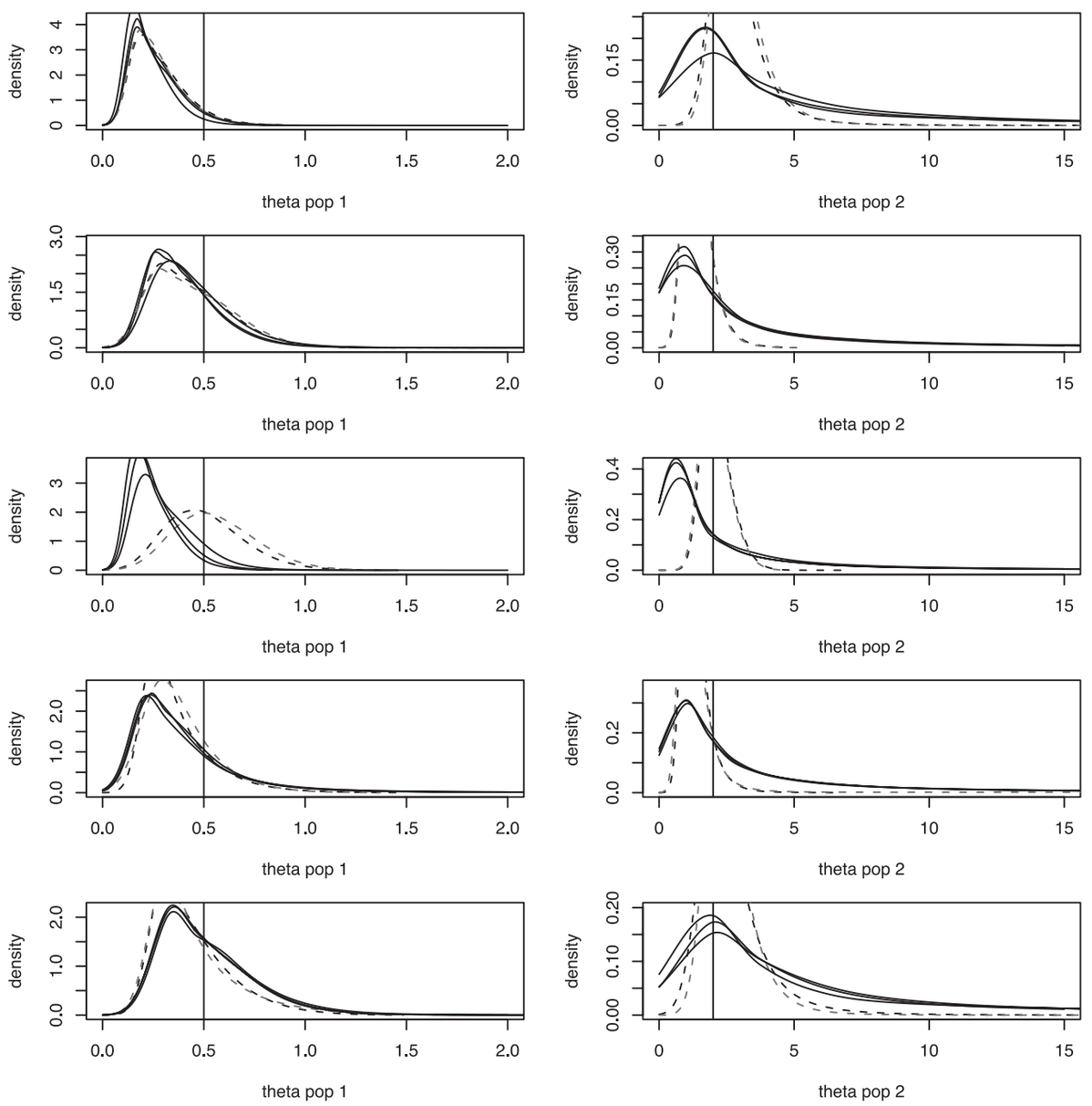

Figure 14.4. *Comparison of the posterior distributions of* Θ_1 *and* Θ_2 *obtained using IM (dashed line) and the ABC method on 5 different data sets.*

The posterior distributions of the six parameters computed for the five different simulated data sets are illustrated in Figures 14.4, 14.5 and 14.6. In general there is quite good agreement between the two approaches. The inferences from IM appeared substantially better only in the case of θ_2, where in four out of five cases the posterior density at the true parameter value was greater for IM than the ABC method. Agreement was best for θ_1 and $\mathcal{M}_2$. Neither method was able to make good inferences about θ_A, $\mathcal{M}_1$, or t. Overall, although caveats about convergence must be borne in mind, it would appear that the ABC method is competitive with a full-data likelihood method, and is substantially faster to run.

3.2. Three populations

A summary of the ability of the method to recover the true tree topology is given in Table 14.1. It can be seen that when divergence times are fixed, very strong inferences are made about the topology. When immi-

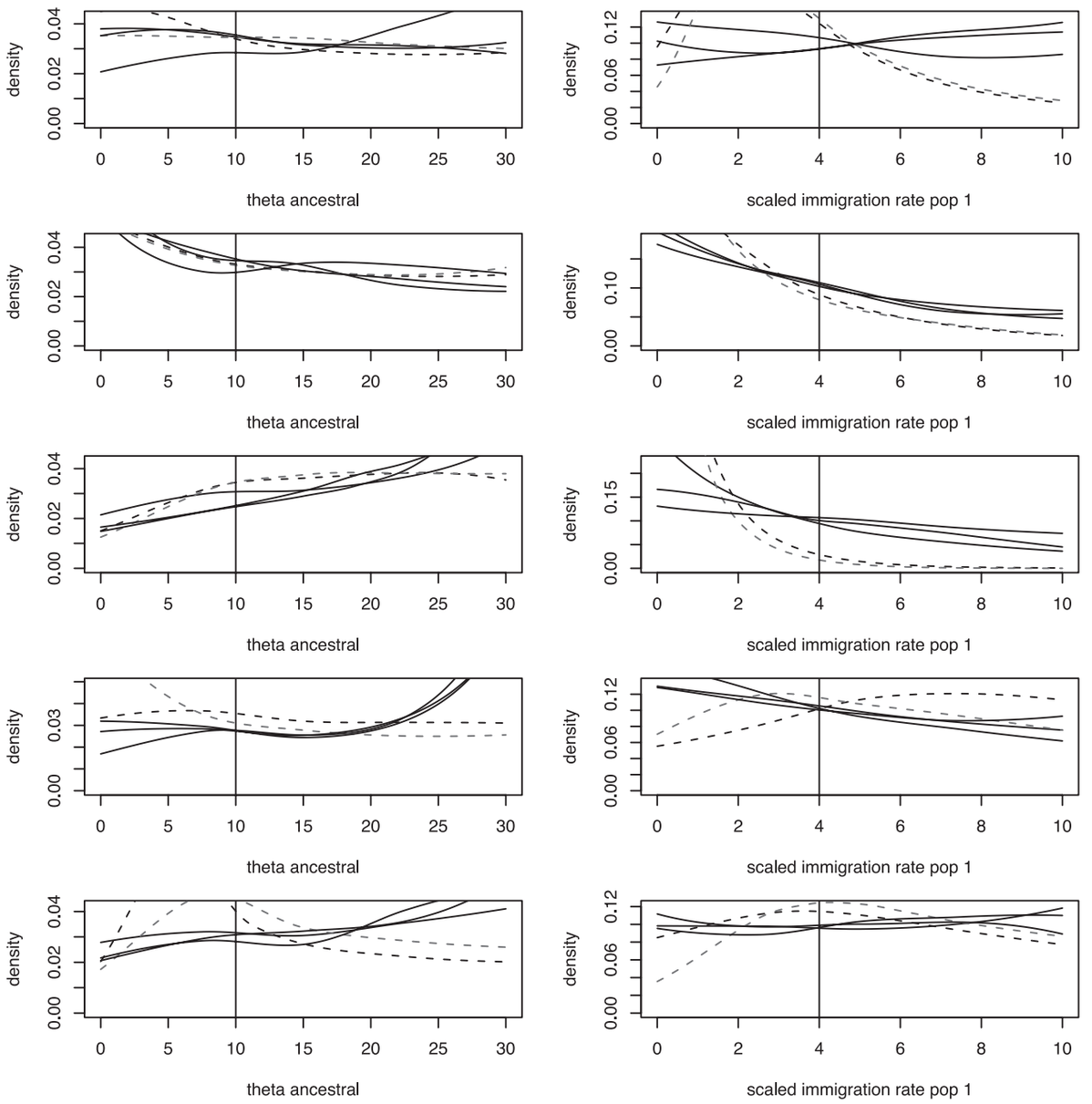

Figure 14.5. *(Continuation of Fig. 14.4.) Comparison of the posterior distributions of* θ_A *and* M_1 *obtained using IM (dashed line) and the ABC method on 5 different data sets.*

gration rates are fixed the mean posterior probability of the true topology is around 0.8. This reduces to only 0.6 when all parameters have wide priors. Increasing the number of loci to 500 only increases the posterior probability to around 0.75. When the posterior probability of the true topology is high, the replicates agree very well, with correlation coefficients near to 1. The fact that the correlation coefficient is only 0.76 when there are wide priors on both immigration and divergence time reflects that the standard error of the logit proportion is proportionately much higher in this case. These standard errors appear to give a good reflection of the reliability of the ABC regression adjustment.

When the other parameters are analysed (Tables 14.2, 14.3 & 14.4) it can be seen that it is in general very difficult to make strong statements about the parameter values, particularly when the prior bounds are as broad as used here. When the

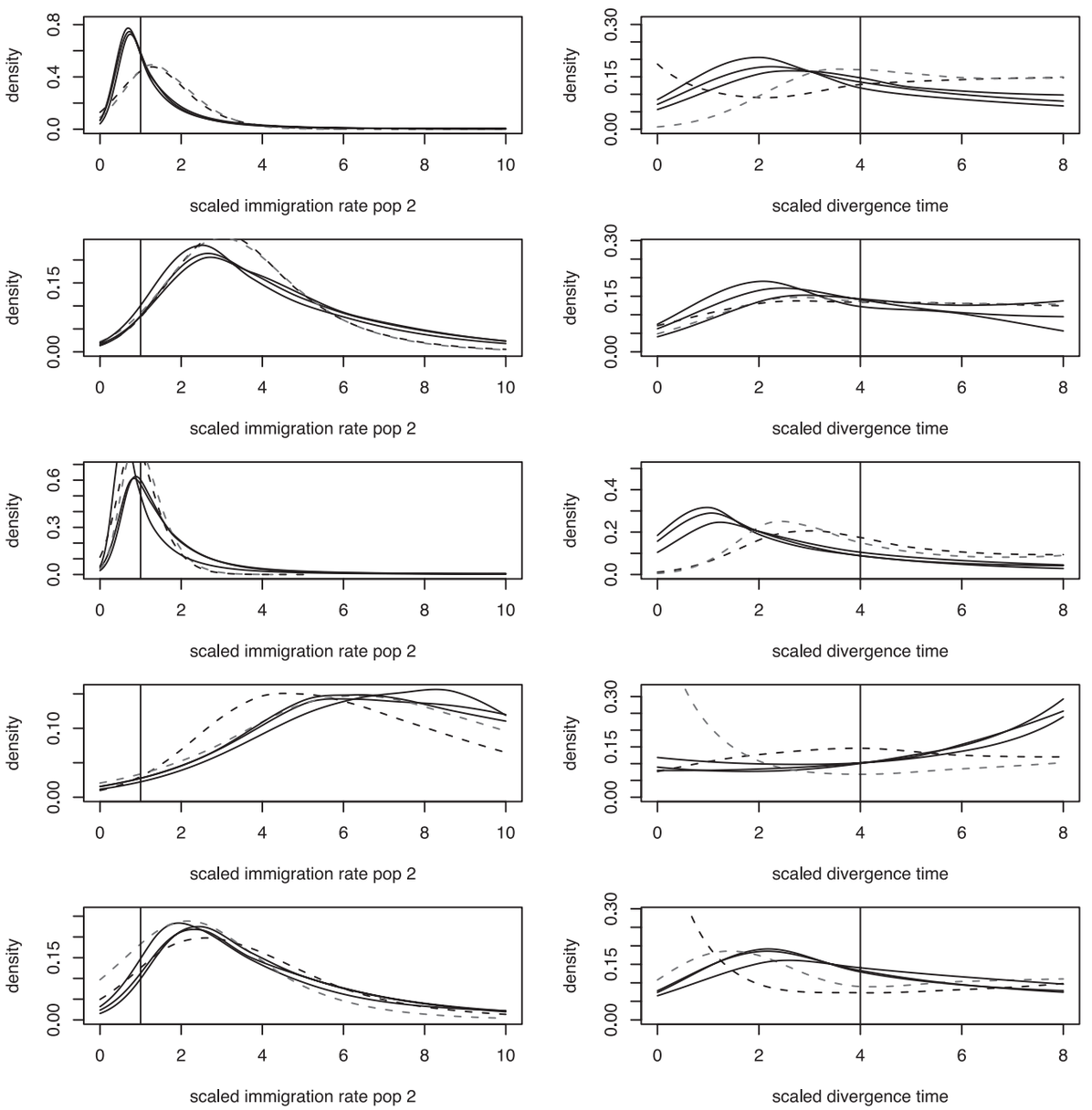

Figure 14.6. *(Continuation of Fig. 14.4.) Comparison of the posterior distributions of $\mathcal{M}_2$ and* t *obtained using IM (dashed line) and the ABC method on 5 different data sets.*

Table 14.1. *Summary of inferences of topology.* $P_{CR}(\cdot)$*: the probability of a topology estimated using categorical regression;* $P_{rej}(\cdot)$*: the probability of a topology estimated using rejection;* $L(\cdot)$*: (log* $P_{CR}(\cdot)$*)/(log* $P_{CR}(3)$*); SE: standard error; cor: Pearman's correlation coefficient between the two replicates in estimates of* $L(1)$*. Other than the correlation coefficient, the results presented are means over the 100 simulated data sets. All results are rounded to two digit precision.*

			$P_{CR}(1)$	$P_{CR}(2)$	$P_{CR}(3)$	$P_{rej}(1)$	$P_{rej}(2)$	$P_{rej}(3)$	$L(1)$	$L(2)$	SE $L(1)$	SE $L(2)$
Fixed T	cor	1										
	rep1		1	2.4e-06	2.9e-05	0.35	0.32	0.32	18	–1.5	0.68	0.77
	rep2		1	4.4e-06	3.6e-05	0.35	0.32	0.33	18	–1.3	0.67	0.76
Fixed m	cor	1										
	rep1		0.81	0.10	0.089	0.34	0.33	0.33	2.3	0.16	0.15	0.15
	rep2		0.82	0.094	0.082	0.33	0.33	0.33	2.4	0.16	0.15	0.15
Free	cor	0.76										
	rep1		0.6	0.19	0.21	0.33	0.34	0.33	1.1	–0.10	0.16	0.16
	rep2		0.57	0.22	0.21	0.32	0.33	0.34	1.0	0.065	0.15	0.16
Free 500	cor	0.95										
	rep1		0.74	0.12	0.15	0.32	0.33	0.35	1.6	–0.21	0.19	0.20
	rep2		0.72	0.11	0.17	0.32	0.33	0.34	1.4	–0.47	0.19	0.20

Table 14.2. *Summary of inferences of demographic parameters with splitting times fixed at true values. Definitions of parameters given in the text. cor: Pearson correlation coefficient of estimated modes in the two replicates; RRMSE: square root of the relative mean square error; lo, hi: respectively lower and upper 0.95 HPD intervals; cov: proportion of simulations in which the true value is found within the HPD intervals. Other than the correlation coefficient, the results presented are means over the 100 simulated data sets. All results are rounded to two digit precision. Note that for the summaries for the prior, when the prior is uniform there is no mode (and corresponding RRMSE), and, in this case, the bounds of the uniform distribution are given instead of the HPD interval.*

		cor	mean	mode	RRMSE mean	RRMSE mode	lo	hi	cov
N_1	rep1	1	1863	731	1.1	0.44	82	5773	1
	rep2		1946	758	1.1	0.43	93	5983	1
	prior		5000		4		0	10,000	
N_2	rep1	1	1900	741	1.1	0.42	92	5787	1
	rep2		1839	699	1.0	0.43	81	5674	1
	prior		5000		4		0	10,000	
N_3	rep1	1	514	121	0.55	0.88	0.17	1834	0.79
	rep2		519	121	0.55	0.88	0.17	1843	0.78
	prior		5000		4		0	10,000	
N_{A1}	rep1	1	2746	1165	1.9	0.64	119	7640	0.99
	rep2		2766	1183	1.9	0.66	124	7699	0.97
	prior		5000		4		0	10,000	
N_{A2}	rep1	1	1175	793	0.56	0.59	0	2666	1
	rep2		1156	769	0.55	0.59	0	2626	1
	prior		5000		4		0	10,000	
m_1	rep1	0.94	0.0017	8.7e-05	2.6	1.0	0	0.0044	1
	rep2		0.0017	7.1e-05	2.5	1	0	0.0043	1
	prior		0.0025		4		0	0.005	
m_2	rep1	0.76	0.0018	3.0e-05	2.7	1.1	0	0.0044	1
	rep2		0.0019	9.1e-05	2.8	1.1	2.7e-06	0.0045	1
	prior		0.0025		4		0	0.005	
m_3	rep1	0.74	0.00078	1.5e-05	0.92	0.97	0	0.0028	0.99
	rep2		0.00078	3.3e-06	0.93	1	0	0.0028	0.99
	prior		0.0025		4		0	0.005	
m_a	rep1	1	0.0016	5e-05	2.4	1.3	0	0.0042	1
	rep2		0.0016	5e-05	2.4	1.3	0	0.0041	1
	prior		0.0025		4		0	0.005	
N_1m_1	rep1	1	2.4	1.1	4.1	1.4	0	7	1
	rep2		2.3	1.0	4.1	1.4	0	7	1
	prior		13	0	24	1	0	35	
N_2m_2	rep1	1	2.5	1.2	4.3	1.6	0	7.3	1
	rep2		2.6	1.2	4.5	1.6	0	7.6	1
	prior		13	0	24	1	0	35	
N_3m_3	rep1	1	0.28	0.11	0.58	0.8	0	0.83	0.64
	rep2		0.28	0.10	0.58	0.8	0	0.83	0.64
	prior		13	0	24	1	0	35	
$N_{A1}m_a$	rep1	1	4.3	0.61	8.4	0.74	0	17	1
	rep2		4.3	0.54	8.4	0.65	0	17	1
	prior		13	0	24	1	0	35	

divergence times are fixed (Table 14.2) the square root of the relative mean square error (RRMSE) is typically not substantially above 1 for modes, but somewhat higher for means. In general the means tend to overestimate parameter values, and modes tend to underestimate them. The confidence interval is given by the 0.95 highest posterior density (HPD) limits. These enclose a region within which the parameter value is found with probability 0.95, and in which the probability density is always greater than or equal to the probability density at the limits. The coverage (proportion of simulations where the true value is within the chosen confidence interval), based on 0.95 HPD intervals, is generally quite good, apart from that for N_3 and N_3m_3. However, the HPD intervals for immigration rates are not substantially different from the prior bounds (or HPD intervals in the case of scaled immigration rates). The correlation in modal estimates between the two replicates are generally good. In the case of immigration rates they are lower. However, here the posterior distributions are almost rectangular, following the priors, and modes would be poorly estimated by any sampling method. Estimates based on the means are generally much better correlated, but the modes have been chosen to illustrate the repeatability of the procedure because they are more sensitive to the shape of the posterior distributions.

Table 14.3. *Summary of inferences of demographic parameters with immigration rates fixed at true values. Details as for Table 14.2. An estimate of the correlation coefficient when a standard deviation is 0 is given as NA.*

		cor	mean	mode	RRMSE mean	RRMSE mode	lo	hi	cov
T_1	rep1	1	527	70	4.4	0.39	0	1898	1
	rep2		517	68	4.3	0.41	0	1863	1
	prior		5000		49		0	10,000	
T_2	rep1	1	2682	1023	1.5	0.21	0	7001	1
	rep2		2644	1000	1.5	0.22	0	6893	1
	prior		10,000	10,000	8.1	8.1	2236	17,764	
N_1	rep1	1	2519	1897	1.6	1	547	5051	0.98
	rep2		2578	1946	1.6	1.0	588	5123	0.97
	prior		5000		4		0	10,000	
N_2	rep1	0.99	2452	1903	1.5	1	536	4988	0.99
	rep2		2482	1892	1.5	1	561	4988	0.98
	prior		5000		4		0	10,000	
N_3	rep1	1	1735	1160	0.83	0.35	295	3758	1
	rep2		1775	1184	0.87	0.37	305	3909	1
	prior		5000		4		0	10,000	
N_{A1}	rep1	1	3468	313	2.5	1.6	0	6841	1
	rep2		3415	444	2.5	1.5	0	6980	1
	prior		5000		4		0	10,000	
N_{A2}	rep1	NA	2231	0	1.3	1	0	7655	1
	rep2		2242	0	1.3	1	0	7620	1
	prior		5000		4		0	10,000	

When the immigration rates are fixed (Table 14.3), the point estimates of divergence times based on the mode are quite good, although the HPD limits tend to be wide. Estimates of population sizes tend to be worse than in Table 14.2, and the modal estimates are biased high for current populations, and low for ancestral populations, with wide HPD limits. The correlation coefficient of modal estimates is generally very good.

When all parameters (other than mutation rate) are free to vary, the quality of inferences falls. The correlation between replicates in estimates of the mode are often quite low, but in these cases, as noted above, the posterior distributions are very similar to the flat priors, and modes would be difficult to estimate under any method, and more likely reflect problems in density estimation than problems in the regression/rejection procedure itself. Relevant to this argument is the observation of a very high correlation in the case of the scaled immigration rates and T_2, which have priors that have a mode — essentially, the posterior distribution follows the prior. The results for T_1, where the correlation is low, and the relatively small bias of the modal estimate appears at variance with the very high RRMSE is explained by the observation that the mode is generally estimated at 0, with respectively 2 and 5 cases out of 100 where the mode is at 10,000 in the two replicates. The population sizes and scaled immigration rates are generally the best-estimated of the parameters. Coverage is poor for T_2 and immigration rates. It is particularly bad for the latter parameters, and may reflect poor behaviour of the regression adjustment, but, again, it should be noted that the posterior distributions tend to be very broad, and these results may also reflect problems in density estimation. With 500 loci the inferences are similar to those from 50 loci, with somewhat narrower HPD limits and lower RRMSEs (results not shown). In particular the coverage for T_2 is similar to that for the 50 locus case, but for m_1 and m_2 it is zero, with the mean lower HPD limit at around 0.001, and never lower than the true value of 0.0005.

Overall, although a much wider set of scenarios needs to be investigated, and conclusions are necessarily tentative, it would appear that there is little scope, without more informative priors, for making strong statements about demographic history using even large numbers of microsatellite loci. Other analyses (not reported here) based on sets of 100 simulated data sets, as here, but under different scenarios (e.g. no immigration, non-identical parameter values) tend to back up this conclusion. Current population sizes and N_{A2} are quite well estimated, but once immigration is included in the prior the population tree topology is very difficult to infer with certainty.

3.3. Analysis of Channel Island fox data

As an example application of the ABC approach, part of a data set on Californian Channel Island foxes has been used, previously analysed in Goldstein *et al.* (1999). The populations chosen for study were those from mainland California (M) (n = 15), the island of Santa Cruz (SCR) (n = 29), and the island of San Clemente (SCL) (n = 30). Nineteen microsatellite loci

Table 14.4. *Summary of inferences of demographic parameters when there are no point priors on times or immigration rates. Details as for Tables 14.2 and 14.3.*

		cor	mean	mode	RRMSE mean	RRMSE mode	lo	hi	cov
T_1	rep1	0.44	3749	100	37	10	0	8248	1
	rep2		3954	500	39	22	0	7768	1
	prior		5000		49		0	10,000	
T_2	rep1	0.95	9000	8657	7.2	7	1016	16,600	0.6
	rep2		9256	9473	7.4	7.7	1170	16,985	0.43
	prior		10,000	10,000	8.1	8.1	2236	17,764	
N_1	rep1	1	1155	543	0.38	0.5	111	3015	1
	rep2		1168	545	0.39	0.5	123	3040	1
	prior		5000		4		0	10,000	
N_2	rep1	0.98	1102	529	0.32	0.51	104	2824	1
	rep2		1095	538	0.32	0.5	98	2862	1
	prior		5000		4		0	10,000	
N_3	rep1	1	1036	459	0.34	0.58	74	2878	1
	rep2		967	429	0.33	0.6	69	2631	1
	prior		5000		4		0	10,000	
N_{A1}	rep1	0.31	3588	102	2.6	1.3	0	7633	1
	rep2		4059	1015	3.1	3	0	6918	1
	prior		5000		4		0	10,000	
N_{A2}	rep1	NA	4375	204	3.4	1.6	0	8815	1
	rep2		4214	0	3.2	1	0	9152	1
	prior		5000		4		0	10,000	
m_1	rep1	0.9	0.0031	0.0043	5.2	8	0.00065	0.005	0.24
	rep2		0.0032	0.0045	5.4	8.3	0.00076	0.005	0.12
	prior		0.0025		4		0	0.005	
m_2	rep1	0.67	0.003	0.0041	5.1	7.5	0.00064	0.005	0.28
	rep2		0.0032	0.0046	5.4	8.4	0.00074	0.005	0.12
	prior		0.0025		4		0	0.005	
m_3	rep1	0.97	0.00083	0.00024	0.99	0.6	0	0.0028	0.98
	rep2		0.00082	2e-04	1	0.67	0	0.0028	0.98
	prior		0.0025		4		0	0.005	
m_a	rep1	0.4	0.0027	0.0047	4.3	8.7	1.6e-05	0.0011	0.62
	rep2		0.0027	0.0049	4.4	9	0.00029	0.0047	0.99
	prior		0.0025		4		0	0.005	
N_1m_1	rep1	0.99	3	1.6	5.4	2.5	0	7.7	1
	rep2		3.1	1.7	5.6	2.7	0	8	1
	prior		13	0	24	1	0	35	
N_2m_2	rep1	0.99	2.9	1.5	5	2.2	0	7.4	1
	rep2		3	1.7	5.3	2.6	0	7.7	1
	prior		13	0	24	1	0	35	
N_3m_3	rep1	1	0.7	0.33	0.79	0.48	0	1.8	0.98
	rep2		0.65	0.32	0.72	0.49	0	1.7	0.97
	prior		13	0	24	1	0	35	
$N_{A1}m_a$	rep1	NA	9.8	0	19	1	0	33	1
	rep2		11	0	22	1	0	36	1
	prior		13	0	24	1	0	35	

were scored. In the analysis by Goldstein *et al.* (1999) the UPGMA consensus tree based on $(\delta\mu)^2$ suggested a grouping of ((SCL, SCR), M), but the bootstrap support given for each clade was around 50%, consistent with a probability of each possible tree topology of around (1/3, 1/3, 1/3). The aim of the analysis was to try to resolve this tree better, and to obtain parameter estimates.

A point prior of 1.5 years was chosen for the generation time. The priors for the divergence times were based on information in Goldstein *et al.* (1999). There was much uncertainty about value of N_e, and I chose generalized gamma distributions with thick upper tails and the bulk of the density towards values <2000, with a lower limit of 20 for the island populations and the first ancestral population and 200 for the mainland and the common ancestral population. A Beta (1, 1000) prior was used for immigration rate in all populations. Densities and summaries of these priors are listed in Figures 14.7, 14.8, and Table 14.5, along with the results.

It can be seen that the effective size of the mainland population is much higher than suggested by the priors. This probably also pulls the mutation rate

towards high values, but this remains to be tested. The effective size of the island populations are smaller than suggested by the priors. There appears to be little or no information in the data on the splitting times (marginal to topology). The point estimate for the probability of ((M, SCR), SCL) is 0.23, that of ((M, SCL), SCR) is 0.26, and that of ((SCR, SCL), M) is 0.52. The log of the ratio of the first two probabilities over the last is respectively –0.822 (0.4 s.e.) and –0.701 (0.41 s.e.). Thus there is slightly stronger evidence in favour of the Mainland population being the outgroup than in the analysis of Goldstein *et al.* (1999), but there is no strong support for any particular topology, and there is some error in these estimates. In addition it should be noted that quite informative priors have been used. The priors on the splitting times were based on the assumption of a ((SCR, SCL), M) topology, which will have some influence on the inferred topology. The immigration rate into the mainland population is inferred to be much lower than suggested by the priors. There is a fair amount of support for zero immigration rates in all the populations (the 0.05 HPD limits include 0 in all cases). Further analyses need to be carried out conditioning on zero migration, and examining the influence of the priors.

3.4. Analysis of San, French, and Orcadian microsatellite data

The study by Rosenberg *et al.* (2002) has provided gene frequency information on microsatellite variation at 377 microsatellite loci surveyed from 52 worldwide human populations. This provides a valuable downloadable resource (http://rosenberglab.bioinformatics.med.umich.edu/datasets.html). From this data set I chose to analyse 3 samples: French (29 individuals), San (7), Orcadians (16). In these samples, 329 loci showed perfect repeats, and the remainder were discarded.

For effective population sizes a gamma prior was used with shape parameter 1.5. In the case of the French and San populations a scale parameter of 20,000 was used, 1000 was used for Orcadians, and 50,000 for N_{A1} and N_{A2}. For the current populations 10 was added to the simulated random variables, and for the ancestral populations 100 was added. For the time to the first split (back in time) the prior was also gamma. In this case the scale was 2000 with shape 2, and 1000 was added to the simulated random variables. For $T_2 - T_1$ a gamma prior was used with scale 10,000 and shape 4. For the immigration rate, a prior was put on the number of immigrants (e.g. $N_1 m_1$ etc.). This was a generalized gamma distribution (a = 0; b = 2.0; c = 1.1; k = 0.8). The shapes of these priors are illustrated in Figures 14.9–14.11. The prior for the mutation rate was lognormal with mean on a log 10 scale of –3.5 and standard deviation of 0.2, the prior for was a normal distribution with mode at 0 and standard deviation of 0.1, truncated at 0.

Table 14.5. *Summary of analysis of Channel Island fox data. The posterior modes and 0.95 HPD limits are shown. The equivalent summaries for the priors are shown underneath. These latter were estimated from simulated data, and are not obtained analytically, and therefore there is some minor variation in estimates for identical prior distributions.*

parameter	mode	HPD lo	HPD hi
μ	–3.07 –3.50	–3.37 –4.09	–2.74 –2.92
σ	0.00 0.00	0.00 0.00	0.275 0.196
T_1	10,900 5140	3250 3210	14,200 14,200
T_2	12,600 12,600	11,100 11,000	17,000 16,700
N_1	2670 350	984 200	7420 2770
N_2	41.7 142	13.2 20	146 2000
N_3	34.3 145	8.81 20	96.8 2010
N_{A1}	205 181	0 20	2310 2560
N_{A2}	424 356	102 200	3180 2730
m_1	0.0000984 0	0 0	0.000324 0.0149
m_2	0.000576 0	0 0	0.00462 0.0149
m_3	0.000871 0	0.00 0	0.00518 0.0152
m_a	0 0	0 0	0.0126 0.0147

For the analysis two replicate samples of 500,000 were taken from the joint distribution of priors and summary statistics. The closest 10,000 points were used in the analysis. Inferences about the topology were very strong. From the categorical regression method the estimated log odds ratios of probability of a ((French, Orcadian), San) and a ((French, San), Orcadian) topology versus a ((San, Orcadian), French) topology were respectively 20.6 (s.e.: 0.64) and –5.7 (0.84) for the first replicate and 19.5 (0.60) –5.2 (0.77) for the second. These translate to estimated posterior probabilities for the 3 topologies (in the order given above) of respectively, (~ 1, 3.8×10^{-12}, 1.1×10^{-9}) and (~ 1, 1.6×10^{-11}, 3.0×10^{-9}). For comparison, the estimates that come from the rejection method are (0.35 0.30 0.35) in both replicates. The difference here (and in the test simulations above) may seem surprising, but it is worth noting that in the first replicate the 10 'nearest' simulated points (in terms of Euclidean distance) all have the ((French, Orcadian), San) topology and in the second replicate this figure is 25.

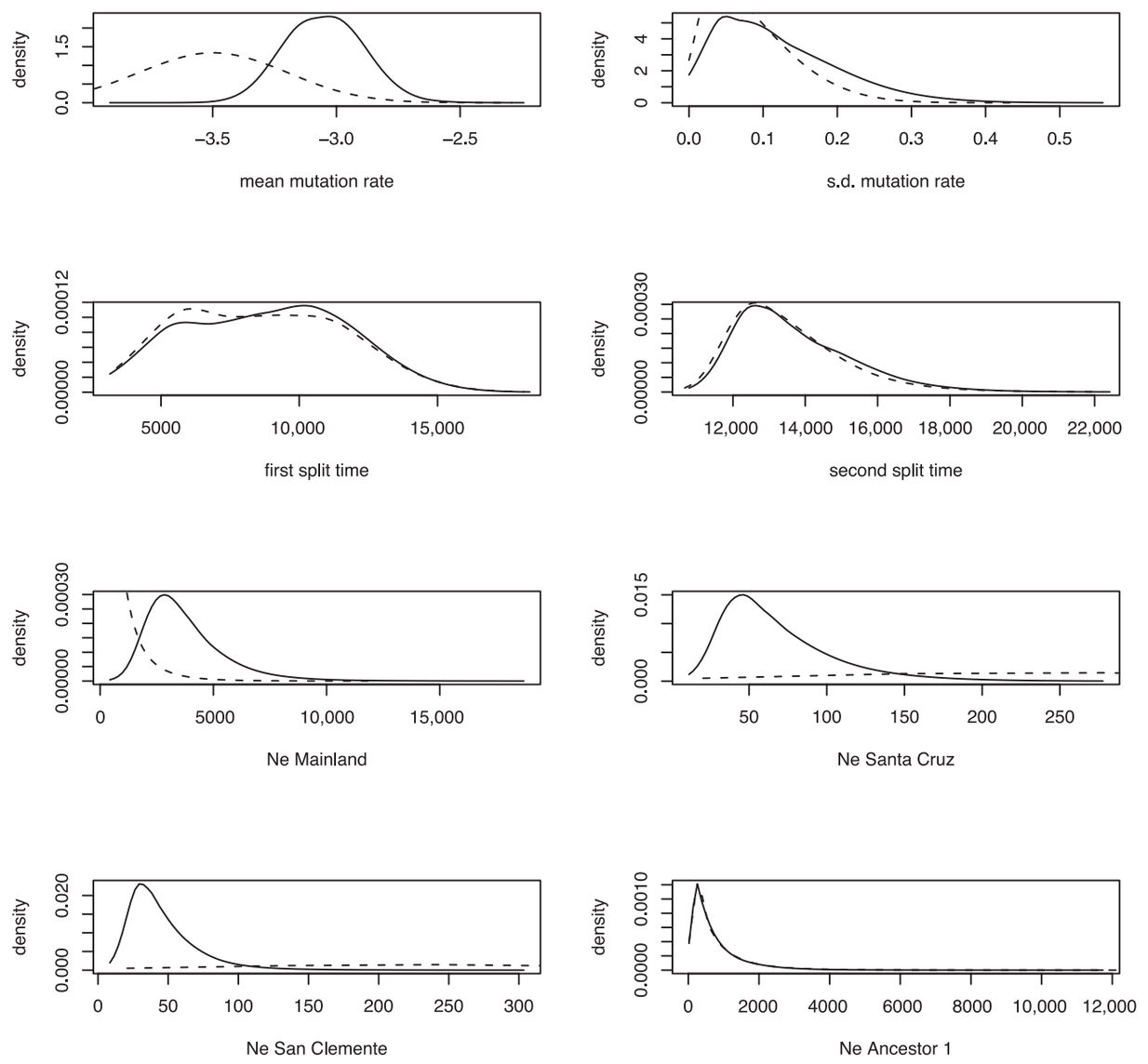

Figure 14.7. *Analysis of Channel Island fox data. Plots of the posterior densities for all the parameters in the model. The prior distribution is shown as a dotted curve. See text for details.*

Inferences about the other parameters are relatively weak. Of the population sizes (Fig. 14.9), the posterior estimates are typically around the 10,000 figure observed in many studies (Harpending *et al.* 1998), although with mutation rate prior used, the mode is somewhat lower than this. There is no strong evidence of any major demographic changes since the populations split. The estimates for the San, and the common ancestor are the most strongly distinguishable from the priors. The modal estimate for the San is around 4000 and around 6000 for the common ancestor.

There appears to be almost no information (in addition to that provided by the priors) on divergence times (Fig. 14.10). The divergence time for the most recent common ancestor is increased somewhat relative to the prior. The immigration rates follow the prior (Fig. 14.10). The estimates for the scaled immigration rates suggest (Fig. 14.11) that immigration into the San is reduced relative to the prior, and increased in the case of the Orcadians, with a mode at around 2 immigrants per generation.

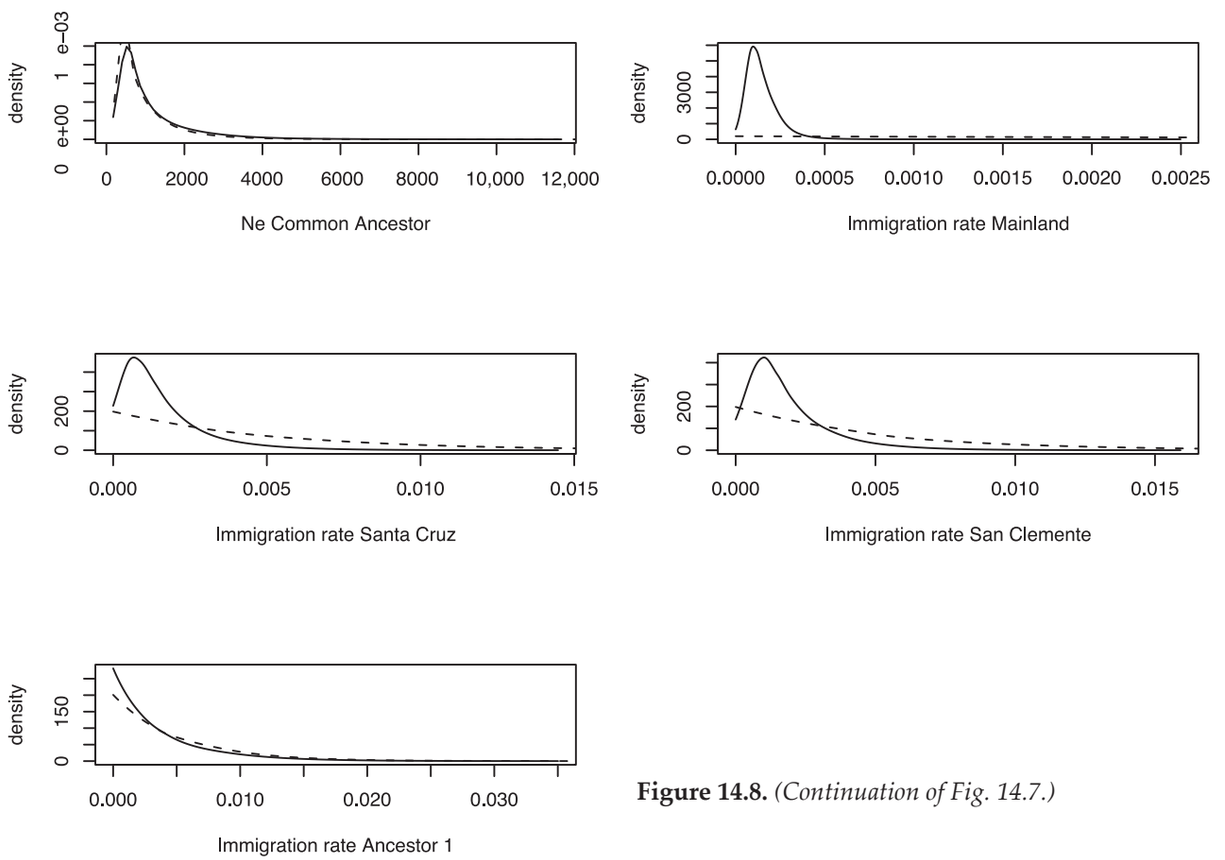

Figure 14.8. *(Continuation of Fig. 14.7.)*

Thus overall, although, reassuringly, the inferred splitting topology strongly supports expectations, there appears to be relatively little information from these 329 loci about the demographic details. Part of the variance in the marginal posterior distributions stems from the assumed uncertainty in mean mutation rate among loci. Typically in most population genetic inference a point prior is used on the mutation rate. The prior used here gives the bulk of support to mean mutation rates in the region of 0.0001 to 0.0006, and it seems doubtful, given current knowledge, and especially given the uncertainty in the correct mutational model for microsatellites (Cornuet *et al.* 2006), whether we can assume any stronger precision.

4. Discussion

With the development of increasingly powerful computational statistical machinery there is a temptation to tackle models of ever increasing complexity. The results presented here seem to suggest that when more complex demographic models are considered our ability to use genetic data to make strong statements about past population history is quite limited. With 19 loci, as with the Channel Island foxes, there is little information to determine the history of population splitting, and with 329 loci in the human data, although the population tree seems to be accurately resolved, there is great uncertainty on the other demographic parameters. Similar points have been made before by Takezaki & Nei (1996), in the case of population trees without migration, who suggested that hundreds of loci would be needed for a reasonable chance of correctly inferring population trees with microsatellite markers. Their observations were based on moment estimators such as $(\delta\mu)^2$ (Goldstein *et al.* 1995), and it might have been hoped that through the use of a suite of summary statistics describing many more aspects of the data than can be captured in a single statistic, better inferences would have been obtained. It is possible that this reflects a limitation of the particular statistical approach taken here. However, in the case of two population models, in comparisons with a likelihood-method that can use all the data (Hey & Nielsen 2004), the ABC procedure does appear perform well.

With three populations the parameter space is much larger and so it is dangerous to argue that the inferences made using ABC are comparable with

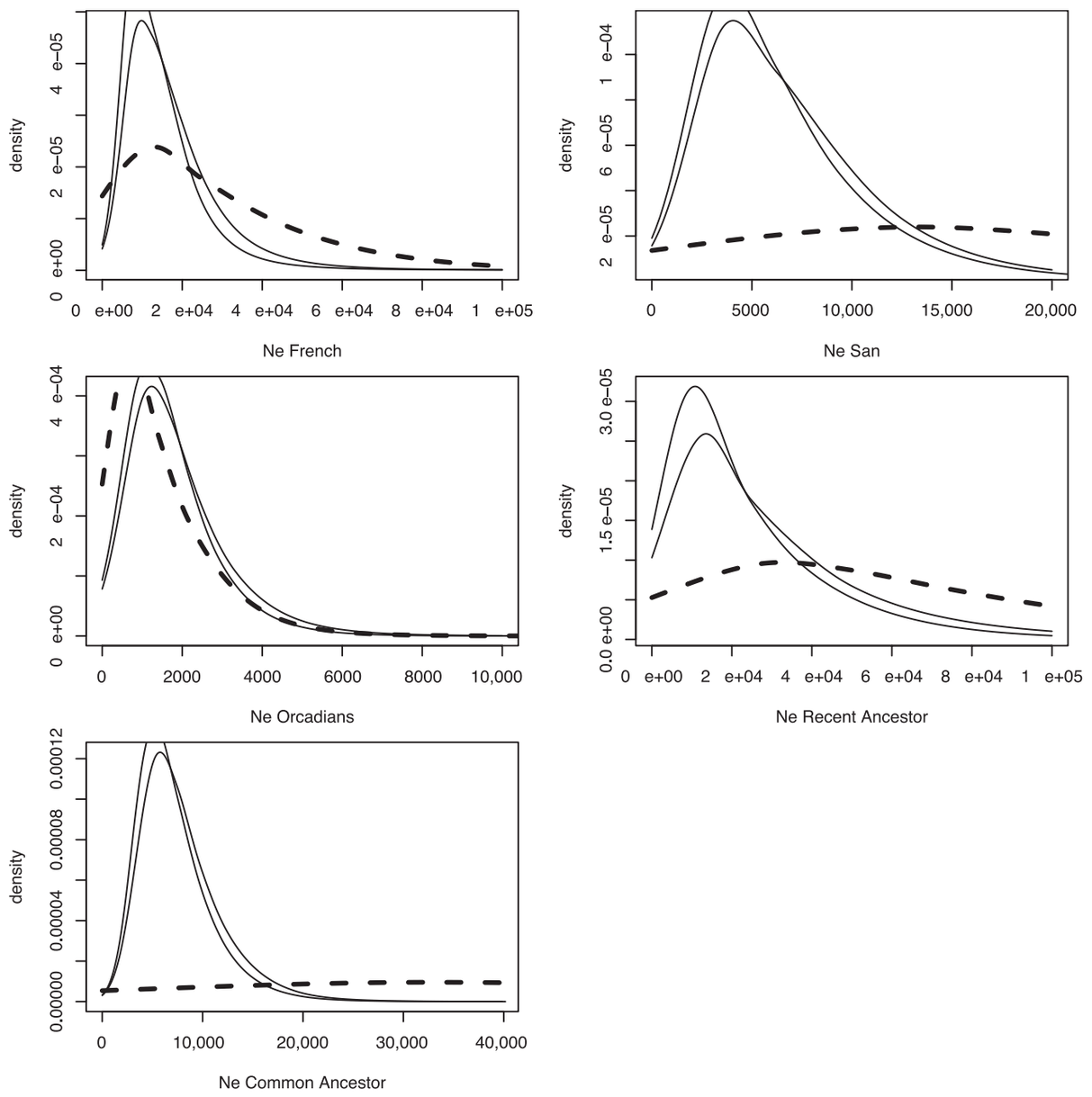

Figure 14.9. *Plots of the posterior densities for effective population sizes in French, San, and Orcadian populations. The prior distribution is shown as a dotted line. The two replicates are shown.*

those that would be made with a MCMC approach on three populations. Much of the problem seems to be caused by the presence of immigration. Simulation tests of the ABC approach without immigration have indicated generally greater accuracy, particularly in the recovery of the correct tree topology (results not shown).

One of the major technical highlights of the results presented here is the substantial improvement in model selection that can be made using categorical regression rather than rejection. Model selection is a relatively challenging area in computational statistics, and typically requires reversible-jump MCMC methods (e.g. Cornuet *et al.* 2006). In principle it is very

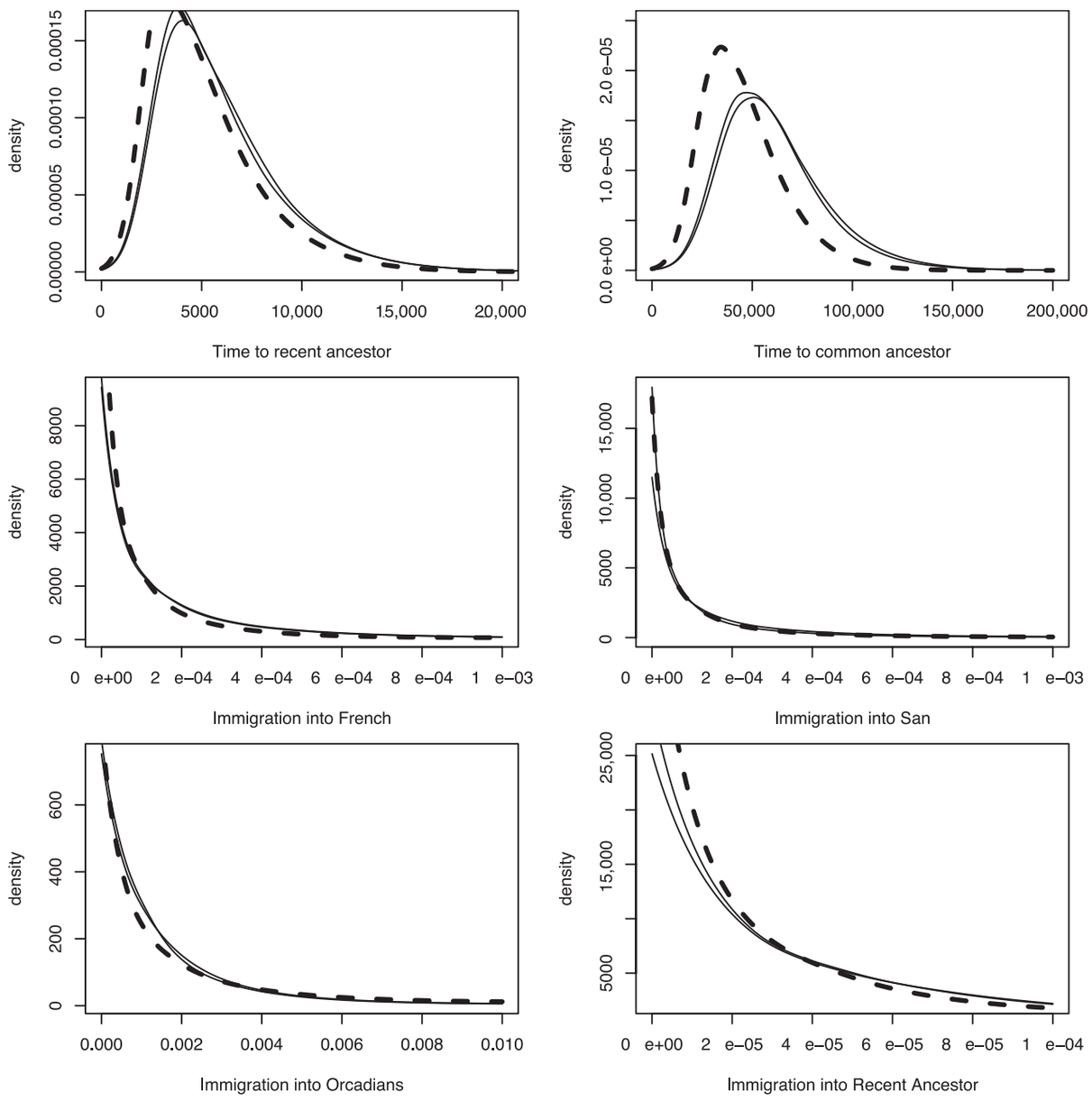

Figure 14.10. *Plots of the posterior densities for immigration rates and times of divergence in French, San, and Orcadian populations.*

easily addressed through the ABC framework (Pritchard *et al.* 1999), but in practice, as illustrated here, the rejection method has relatively little accuracy.

One problem that was noted in the ABC analysis when making comparisons with IM is that the current method of sampling from the prior is inefficient when the posterior distribution is very different from the prior, as with some of the test parameters. This is probably a critical limitation to the ABC approach as described in Beaumont *et al*. (2002), because it means that often the summary statistics measured from the data may lie on the edges of the simulated joint distribution, and accurate conditional density estimation will rely heavily on the linearity of the

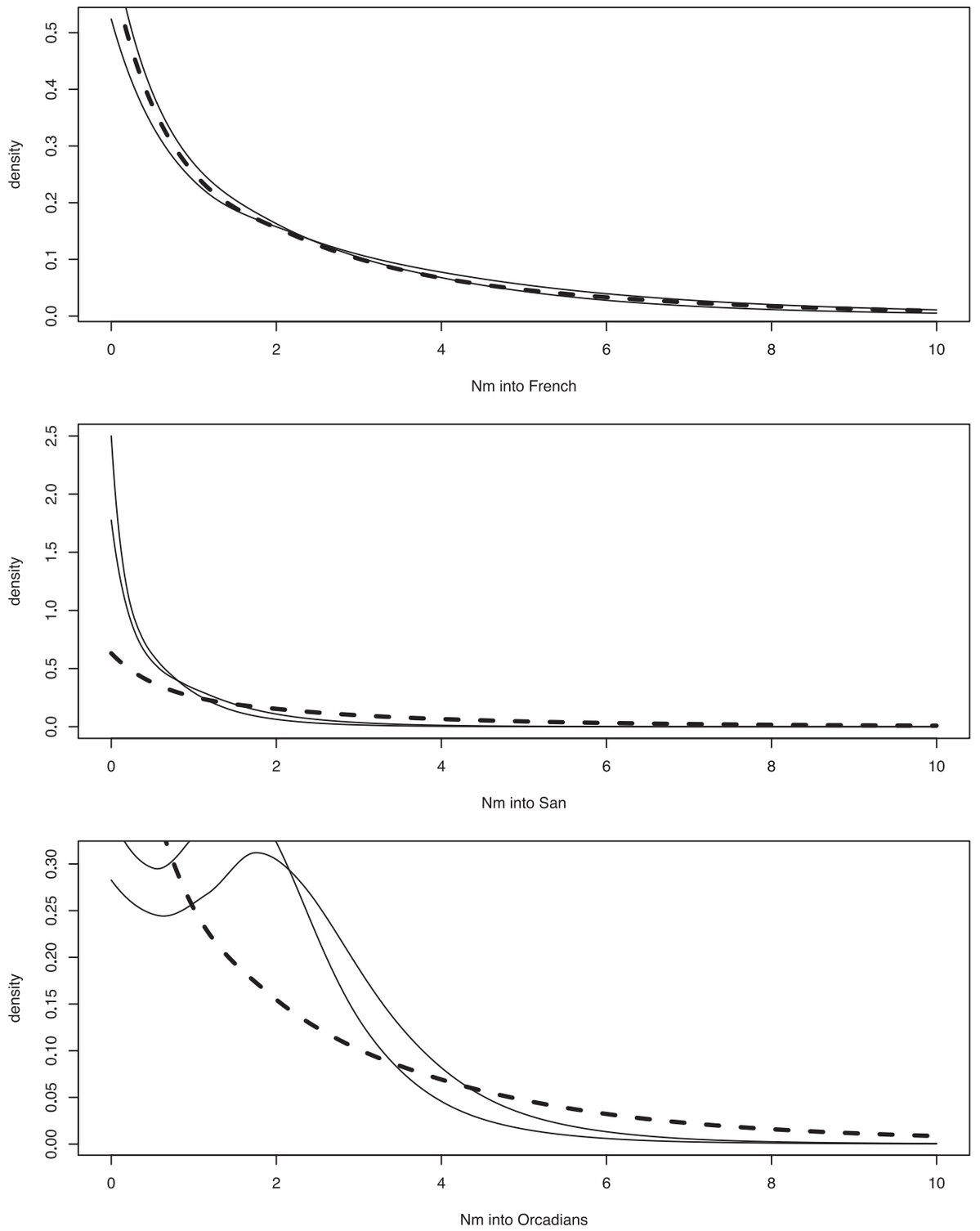

Figure 14.11. *Plots of the posterior densities for scaled immigration rates (number of immigrants per generation) in French, San, and Orcadian populations.*

regression. Errors in the regression adjustment may well explain, for example, the relatively poor performance in comparison with IM for 2 (Fig. 14.4), and the poor coverage noted for some immigration rates in Table 14.4. A more efficient approach might be to sample from a distribution that is closer to the posterior distribution and then reweight the simulated points by the ratio of the probability density of the parameter values under the prior to that under the sampling distribution. An iterative, adaptive, scheme may be useful in this regard.

The analysis of microsatellite data is only one part of the IM program (Hey & Nielsen 2004), which is primarily designed to analyse sequences, and can also work with microsatellites linked to short sequences. Different types of markers can be mixed together in the same analysis. These aspects should not be problematic to implement in an ABC approach, which will also be able to incorporate recombination. The only way to try to overcome the difficulties posed by the relatively low information content of markers such as microsatellites is through the analysis of multilocus nuclear sequence data. Since recombination will need to be included in such a model, it is likely that only approximate methods will be feasible with such data, and this represents a suitable future goal in the development of the ABC approach described here.

5. Appendix

5.1. Regression-based method for conditional density estimation

In this method we assume that we have measured a d dimensional vector of summary statistics s from a data set. We have n random draws of a (scalar) parameter $\Phi_{1,\ldots,n}$ and corresponding summary statistics $S_{1,\ldots,n}$ simulated from the joint distribution of parameters and summary statistics $P(S, \Phi)$. (The model may have any number of parameters, which can be considered jointly, but the regression adjustment described here is applied to one parameter at a time.) We scale s and S so that each summary statistic in S has unit variance.

We use the method of local-linear regression to compute the posterior mean $E(\Phi \mid S = s)$ (see, for example, Ruppert & Wand (1994), for background to the approach). In this method we want to minimize

$$\sum_{i=1}^{n} \{\Phi_i - \alpha - \beta^T (S_i - s)\}^2 K_\delta(\| S_i - s \|)$$

where

$$\alpha = E(\Phi \mid S = s),$$
$$\| x \| = \sqrt{\sum_{i=1}^{d} x_i^2},$$

and we use the Epanechnikov kernel

$$K_\delta(t) = \begin{cases} c\delta^{-1}\left(1 - (t/\delta)^2\right) & t \le \delta \\ 0 & t > \delta \,. \end{cases}$$

The solution is

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} = (S_s^T W_s S_s)^{-1} S_s^T W_s \Phi$$

where

$$\Phi = [\Phi_1, \cdots, \Phi_n]^T$$

$$W_s = \mathrm{diag}\{K_\delta(\| S_1 - s \|), \cdots, K_\delta(\| S_n - s \|)\}$$

$$S_s = \begin{bmatrix} 1 & (S_1 - s)^T \\ \vdots & \vdots \\ 1 & (S_n - s)^T \end{bmatrix}$$

Our best estimate of the posterior mean is then

$$\hat{\alpha} = e_1^T (S_s^T W_s S_s)^{-1} S_s^T W_s \Phi$$

where e_1 is a $d + 1$ length vector $(1, 0, \cdots, 0)$.

In order to estimate posterior densities, Beaumont *et al.* (2002), took a heuristic approach, in which they make an assumption that the errors are constant in the interval and adjust the parameter values as

$$\Phi_i^* = \Phi_i - (S_i - s)^T \hat{\beta}.$$

The posterior density for can be approximated as

$$\widehat{\pi}(\Phi \mid S = s) = \frac{\sum_i K_\Delta(\Phi_i - \Phi) K_\delta(\| S_i - s \|)}{\sum_i K_\delta(\| S_i - s \|)}$$

where $K_\Delta(t)$ is another Epanechnikov kernel with bandwidth Δ. Alternatively some other density method can be used, and in this paper the local-likelihood method of Loader (1996) is used, implemented in Locfit under R, weighting the points with $K_\delta(\| S_i - s \|)$ as above. In Beaumont *et al.* (2002) the 'tolerance' of the method was not measured directly in terms of the Epanechnikov bandwidth , but in terms of P_δ, the proportion of simulated points where $\| S_i - s \| \le \delta$.

Note

1. The final revision of this chapter was submitted 18/02/06. No attempt has been made to update at the proof stage the many developments in this area that have occurred in the last two years.

References

Beaumont, M.A. & R.A. Nichols, 1996. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London*, Series B 263, 1619–26.

Beaumont, M.A., W. Zhang & D.J. Balding, 2002. Approximate Bayesian computation in population genetics. *Genetics* 162, 2025–35.

Cavalli-Sforza, L.L. & A.W.F. Edwards, 1967. Phylogenetic analysis: models and estimation procedures. *Evolution* 32, 550–70.

Cornuet, J.M., M.A. Beaumont, A. Estoup & M. Solignac, 2006. Inference on microsatellite mutation processes in the invasive mite, Varroa destructor, using reversible jump Markov chain Monte Carlo. *Theoretical Population Biology* 69, 129–44.

Excoffier, L., A. Estoup & J.M. Cornuet, 2005. Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics* 169, 1727–38.

Goldstein, D.B., A. Ruiz Linares, L.L. Cavalli-Sforza & M.W. Feldman, 1995. Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academy of Sciences of the USA* 92, 6723–7.

Goldstein, D.B., G.W. Roemer, D.A. Smith, D.E. Reich, A. Bergman & 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.

Green, P.J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–32.

Griffiths, R.C. & S. Tavaré, 1994. Simulating probability distributions in the coalescent. *Theoretical Population Biology* 46, 131–59.

Harpending, C., M.A. Batzer, M. Gurven, L.B. Jorde, A.R. Rogers & S.T. Sherry, 1998. Genetic traces of ancient demography. *Proceedings of the National Academy of Sciences of the USA* 95, 1961–7.

Hey, J. & R. Nielsen, 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* 167, 747–60.

Hudson, R.R., 2001. Two-locus sampling distribution and their application. *Genetics* 159, 1805–17.

Jennings, W.B. & S.V. Edwards, 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution* 59, 2033–47.

King, J.P., M. Kimmel & R. Chakraborty, 2000. A power analysis of microsatellite based statistics for inferring past population growth. *Molecular Biology and Evolution* 17, 1859–68.

Kuhner, M.K., J. Yamato & J. Felsenstein, 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* 140, 1421–30.

Li, N. & M. Stephens, 2003. Modelling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–33.

Loader, C.R., 1996. Local likelihood density estimation. *Annals of Statistics* 24, 1602–18.

Marjoram, P., J. Molitor, V. Plagnol & S. Tavaré, 2003. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the USA* 100, 15,324–8.

McVean, G., P. Awadalla & P. Fearnhead, 2002. A coalescent based method for detecting and estimating recombination from gene sequences. *Genetics* 160, 1231–41.

Nielsen, R. & J. Wakeley, 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics* 158, 885–96.

Pritchard, J.K., M.T. Seielstad, A. Perez-Lezaun & M.W. Feldman, 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution* 16, 1791–8.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber, *et al.*, 2002. Genetic structure of human populations. *Science* 298, 2981–5.

Ruppert, D. & M.P. Wand, 1994. Multivariate locally weighted least squares regression. *Annals of Statistics* 22, 1346–70.

Saitou, N. & M. Nei, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–25.

Storz, J.F. & M.A. Beaumont, 2002. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* 56, 154–66.

Takezaki, N. & M. Nei, 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite data. *Genetics* 144, 389–99.

Wilson, I.J. & D.J. Balding, 1998. Genealogical inference from microsatellite data. *Genetics* 150, 499–510.

Wilson, I.J., M.E. Weale & D.J. Balding, 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society* A 166, 155–88.

Wright, S., 1931. Evolution in Mendelian populations. *Genetics* 16, 97–159.

Wright, S., 1937. The distribution of gene frequencies in populations. *Proceedings of the National Academy of Sciences of the USA* 23, 307–20.