

LIKELIHOOD ANALYSIS OF ONGOING GENE FLOW AND HISTORICAL ASSOCIATION

RASMUS NIELSEN^{1,2} AND MONTGOMERY SLATKIN¹

¹Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Abstract.—We develop a Monte Carlo–based likelihood method for estimating migration rates and population divergence times from data at unlinked loci at which mutation rates are sufficiently low that, in the recent past, the effects of mutation can be ignored. The method is applicable to restriction fragment length polymorphisms (RFLPs) and single nucleotide polymorphisms (SNPs) sampled from a subdivided population. The method produces joint maximum-likelihood estimates of the migration rate and the time of population divergence, both scaled by population size, and provides a framework in which to test either for no ongoing gene flow or for population divergence in the distant past. We show the method performs well and provides reasonably accurate estimates of parameters even when the assumptions under which those estimates are obtained are not completely satisfied. Furthermore, we show that, provided that the number of polymorphic loci is sufficiently large, there is some power to distinguish between ongoing gene flow and historical association as causes of genetic similarity between pairs of populations.

Key words.—Coalescence theory, gene flow, likelihood ratio tests, maximum likelihood, population subdivision, single nucleotide polymorphisms.

Received July 8, 1998. Accepted September 1, 1999.

A variety of methods exist for estimating the migration rate in a subdivided population under the assumption that an equilibrium between gene flow and genetic drift has been achieved at neutral genetic loci. There are also methods for estimating the divergence times of populations under the assumption that there is no ongoing gene flow between them. In fact, Wright's F_{ST} can be used for both purposes because F_{ST} bears a simple relationship to the number of migrants exchanged in an equilibrium island model and to the divergence time of completely isolated populations. However, few methods exist that allow the joint estimation of population divergence times and levels of gene flow, although papers by Wakeley (1996, 1998) and Templeton et al. (1995) provide some results that are applicable to mitochondrial DNA and other nonrecombining genomic regions.

To distinguish between ongoing gene flow and historical association, additional information must be extracted from the data: F_{ST} and other simple statistics are not adequate. A natural way to extract more information from data is to find the likelihood of the data under different assumptions. Beerli and Felsenstein (1999) and Bahlo and Griffiths (2000) have written programs to find the likelihood of a sample under a variety of mutation models for nonrecombining genomic regions in an island model, and Nath and Griffiths (1996) have outlined a general method of likelihood analysis of subdivided populations. The problem is that, unless sample sizes are very small, finding the likelihood of a sample when there is ongoing migration is very time consuming because of the large number of possible configurations of the sample. In fact, it is difficult to obtain even a single estimate of the migration rate in an equilibrium island model under the assumption that the model is correct. Consequently, it is currently impossible in practice to explore a variety of assumptions about the demographic history of populations.

Ignoring the effects of mutation greatly simplifies the cal-

ulation of the likelihood. Nielsen et al. (1998) have already used this simplification for estimating divergence times and population phylogenies when there is no ongoing gene flow. Their method provided a relatively rapid way to calculate likelihoods for sample sizes of up to 100 individuals and 100 polymorphic loci. Ignoring mutation is appropriate when the time of divergence of populations is small. We also assume only two alleles per locus, which is appropriate for restriction fragment length polymorphisms (RFLPs) and single nucleotide polymorphisms (SNPs).

The method described in this paper is a generalization of that of Nielsen et al. (1998). The difference is that here we consider only two populations and allow for ongoing gene flow at a rate m after their divergence. We develop a Monte Carlo method for estimating the likelihood of the data as a function of both the migration rate and the time of divergence. From the likelihood function, we can jointly estimate these two parameters and also can carry out likelihood-ratio tests of the hypotheses that either there is no ongoing gene flow or that the two populations diverged sufficiently long ago that an equilibrium between gene flow and genetic drift has been reached.

We apply our method to simulated data and show that it provides usable results for realistic sample sizes and some power to distinguish between different hypotheses. Our method of parameter estimation and hypothesis testing, like any method based on the calculation of a likelihood, relies on an assumed model. We assume a particular model of population history (populations of constant size and gene flow occurring at a constant rate since divergence) and a particular prior distribution of allele frequencies before divergence (a beta distribution). We show that the assumption of a beta distribution as a prior is not critical, because we find that data simulated under a different prior distribution can still lead to reasonably accurate estimates of the parameter values when a beta prior is assumed. There is less we can say about the demographic model, which is unlikely to be realistic for any natural populations. Results from applying our method

² Present address: Department of Organismic and Evolutionary Biology, Harvard University, 288 Biology Laboratories, Cambridge, Massachusetts 02138; E-mail: rnielsen@oeb.harvard.edu.

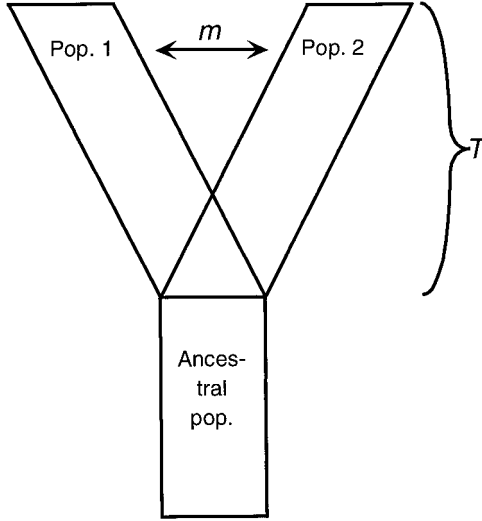


FIG. 1. The demographic model analyzed in this paper. T is the population divergence time and m is the migration rate.

must be interpreted with caution. Obtaining joint estimates of the migration rate and divergence time does not mean that the population from which the data were taken follows the model. Instead it means that the data are comparable to data from an idealized population with those parameter values. A similar problem arises with the application of any other indirect method. The goal of our analysis is to show that a likelihood approach does have some potential to distinguish between ongoing gene flow and historical association as a cause of genetic similarity, something that F_{ST} and similar statistics based on variances of diversity within and between populations are in principle unable to do.

MONTE CARLO METHOD

We assume that two populations, each with N individuals, were derived from an ancestral population at time T in the past, where T is measured in units of $2N$ generations. Since the separation of the two populations, symmetric migration has occurred between the populations at a rate m , per generation, or $M = 4Nm$, in units of $2N$ generations. The populations are of equal and constant size, an assumption that can easily be relaxed. The demographic model is illustrated in Figure 1. All of our results are in terms of the scaled parameters, M and T , which is to be expected when the only force creating differences between populations is genetic drift.

We will here consider data from diallelic loci, so the sample configuration for a single population can be represented as the number of each of two alleles (say A and a) at each locus in the sample. Nielsen et al. (1998) show that this method can be generalized to a model with arbitrary number of alleles per locus.

We will consider the ancestral process of the sample to develop a Monte Carlo method for estimating sample probabilities. The method is analogous to the method developed by Nath and Griffiths (1996) and Nielsen (1998), both of which are based on the Griffiths and Tavaré (1994a,b) simulation algorithm. We will first describe how to calculate the

likelihood for one locus and then describe how the likelihood is computed for multiple independent loci.

Let the current sample configurations in populations 1 and 2 be $\{A_1, a_1\}$, $n_1 = A_1 + a_1$ and $\{A_2, a_2\}$, $n_2 = A_2 + a_2$, where n_i is the sample size obtained from population i , A_i is the number of A alleles, and a_i is the number of a alleles in the sample from population i , ($i = 1, 2$). We assume no mutation, but allow for both genetic drift and migration. In a way analogous to the methods described by Griffiths and Tavaré (1994a,b), Nath and Griffiths (1996), and Nielsen (1998), we find a recursion equation that gives the probability of a sample configuration of the data as a function of all possible ancestral samples at a time just before the last migration or coalescence event. Starting at the present and looking backward in time, one of the following two events could be the most recent event: a migration event or a coalescence event. The rate at which coalescence events occurs in each population is $n_i(n_i - 1)/2$, with $i = 1, 2$, when time is scaled by $2N_e$, where N_e is the inbreeding effective population size of each (diploid) population. The rate at which migration events occur in population i is $Mn_i/2$, with $i = 1, 2$, where $M = 4N_e m$. The total rate of migration and coalescence events is therefore $\lambda = n_1[n_1 - 1 + M]/2 + n_2[n_2 - 1 + M]/2$.

If a coalescence event happened in a lineage carrying allele A in population 1, the ancestral configuration must have been $\{A_1 - 1, a_1, A_2, a_2\}$. From state $\{A_1 - 1, a_1, A_2, a_2\}$, the probability of a transition to the observed sample configuration, given that the last event is a coalescence event in population 1, is $(A_1 - 1)/(n_1 - 1)$. The probability that the last event was a coalescence event in population 1 is $n_1(n_1 - 1)/(2\lambda)$. The total contribution to the probability of the observed sample configuration from this type of event is then:

$$\left[\frac{n_1(n_1 - 1)}{2\lambda} \right] \left[\frac{A_1 - 1}{n_1 - 1} \right] \Pr(\{A_1 - 1, a_1, A_2, a_2\}). \quad (1)$$

The relevant terms can be found similarly for all other possible transitions involving coalescent events.

If the last event was a migration event from population 1 to population 2 of a lineage carrying allele A , the ancestral configuration must have been $\{A_1 + 1, a_1, A_2 - 1, a_2\}$ and the conditional transition probability to the observed configuration is $(A_1 + 1)/(n_1 + 1)$. The probability that the last event was a migration event to population 2 is $Mn_2/(2\lambda)$. The total contribution to the probability of the observed sample configuration from this type of event is then:

$$\left[\frac{Mn_2}{2\lambda} \right] \left[\frac{A_1 + 1}{n_1 + 1} \right] \Pr(\{A_1 + 1, a_1, A_2 - 1, a_2\}). \quad (2)$$

The remaining probabilities for the migration case can be found by similar reasoning. Notice that there are a total of eight possible transitions (four coalescence events and four migration events) that can lead to the observed state in one transition.

If the next event (coalescence or migration) occurs before time t , $t < T$, then we can combine the above expressions to write:

$$\Pr_2(\{A_1, a_1, A_2, a_2\} | \tau < T, M) = \sum_j \lambda_j \Pr_2(\Omega_j | T - \tau, M), \quad (3)$$

where λ_j is the coefficient associated with the j th event

$$\left(\text{e.g., } \lambda_1 = \left[\frac{n_1(n_1 - 1)}{2\lambda} \right] \left[\frac{A_1 - 1}{n_1 - 1} \right] \right),$$

and Ω_j is the ancestral configuration before event j happened, with $j = 1, 2, \dots, 8$. $\Pr_2(\dots)$ is the configuration probability of samples from the two populations. This expression follows directly from the above representation. A similar equation is derived in Nath and Griffiths (1996) for the analysis of equilibrium migration models under a general k -allele model.

If no coalescence or migration events happen before time T in the past, then:

$$\begin{aligned} & \Pr_2(\{A_1, a_1, A_2, a_2\} | \tau \geq T) \\ &= \Pr_1(\{A_1 + A_2, a_1 + a_2\}) \frac{\binom{A_1 + A_2}{A_1} \binom{a_1 + a_2}{a_1}}{\binom{A_1 + A_2 + a_1 + a_2}{A_1 + a_1}}, \quad (4) \end{aligned}$$

where τ is the time of the next coalescence or migration event, and $\Pr_1(\dots)$ denotes the sampling probability of samples taken from a single isolated population, which is ancestral to the two descendent populations.

By conditioning on the time of the next coalescence or migration event, we can now combine the expressions:

$$\begin{aligned} & \Pr_2(A_1, a_1, A_2, a_2 | T, M) \\ &= \int_0^T \lambda e^{-\lambda\tau} \Pr_2(A_1, a_1, A_2, a_2 | \tau < T, M) d\tau \\ &+ e^{-\lambda T} \Pr_1(A_1, a_1, A_2, a_2 | \tau \geq T), \quad (5) \end{aligned}$$

as in Nielsen (1998). This representation of the sampling probability immediately suggests a recursive Monte Carlo integration technique for evaluating the sampling probability. The i th estimate of the probability (p_i) in a Monte Carlo integration scheme can be obtained by the following simulation method. (1) Set $p_i = 1$. (2) Simulate a random exponential variable (t) with rate λ . (3) If $t \geq T$, set

$$p_i = p_i \Pr_1(\{A_1 + A_2, a_1 + a_2\}) \frac{\binom{A_1 + A_2}{A_1} \binom{a_1 + a_2}{a_1}}{\binom{A_1 + A_2 + a_1 + a_2}{A_1 + a_1}}.$$

(4) If $t < T$ and if $\sum_j \lambda_j = 0$, set $p_i = 0$. (5) If $t < T$ and if $\sum_j \lambda_j > 0$, choose an ancestral sample configuration according to the probabilities given by $\lambda_j / \sum_i \lambda_i$. If event j happened, set $p_i = p_i \sum_j \lambda_j$, update the sample to type Ω_j , set $T = T - t$ and repeat the simulation scheme from step 2 with the new sample.

Then $1/k \sum_{i=1}^k p_i \rightarrow \Pr_2(\{A_1, a_1, A_2, a_2\} | T, M)$ when $k \rightarrow \infty$. This simulation scheme is similar to the simulation scheme applied in Nielsen (1998) to estimate divergence times and the simulation scheme applied by Nath and Griffiths (1996) to estimate migration rates.

The above simulation scheme describes how to estimate $\Pr_2(\{A_1, a_1, A_2\} | T, M)$ assuming that there is no bias in the sampling of loci. However, in most studies only variable

markers are included in the analysis. Invariable markers are typically not detected or not reported. To appropriately correct for this sampling bias, it is necessary to condition on variability in the sample. This conditioning is easily achieved by calculating:

$$\begin{aligned} & \Pr_2(\{A_1, a_1, A_2, a_2\} | T, M, \text{variability}) \\ &= \frac{\Pr_2(\{A_1, a_1, A_2, a_2\} | T, M)}{1 - 2\Pr_2(\{A_1 + a_1, 0, A_2 + a_2, 0\} | T, M)}, \quad (6) \end{aligned}$$

assuming that the prior distribution of allele A and allele a is symmetric. Equation (6) provides the likelihood function we will use to estimate M and T . A computer implementation of the algorithm was checked by comparison with analytical results for the case $m = 0$ (e.g., Nielsen et al. 1998) and by comparison with simple coalescence simulation results in cases of $m > 0$.

The above simulation scheme describes how to estimate the likelihood for a single locus. For multiple unlinked loci that are assumed to be in linkage equilibrium, the likelihood for a given set of parameter values is simply obtained by multiplying the likelihoods obtained for different loci. The assumption of unlinked loci is reasonable for most allozyme data, RFLPs, and (SNPs).

An importance sampling scheme for estimating multiple values of M simultaneously was implemented. This is the estimation approach called ‘‘surface simulation’’ in Griffiths and Tavaré (1994a). However, in the present case estimates of the likelihood obtained by this method were found to have very large variances. This method was therefore abandoned and independent estimation of the likelihood for each value of M was used instead.

When applying this method, the prior distribution of allele frequencies, $\Pr_1(\{A, a\})$, needs to be calculated directly or estimated. There are several possibilities. For example, one could try to estimate the ancestral frequencies. In that case, the frequency (f) of one of the alleles, say allele A , in the ancestral population, would be a parameter of the model. $\Pr_1(\{A, a\})$ would then be given by the binomial sampling probability:

$$\Pr_1(\{A, a\} | f) = \binom{A + a}{A} f^A (1 - f)^a, \quad (7)$$

where A and a are the counts of A and a alleles in the sample. The disadvantage of this approach is that the number of parameters will increase linearly with the number of loci included in the analysis. Alternatively, one could assume a prior distribution for the allele frequencies in the ancestral population. For example, one could assume no prior knowledge of the allele frequencies and let them follow a uniform distribution. Then the single population sample probabilities can be obtained by integrating a binomial over a uniform distribution:

$$\begin{aligned} \Pr_1(\{A, a\} | f) &= \int_0^1 1 \cdot \binom{A + a}{A} f^A (1 - f)^a df \\ &= \frac{1}{A + a + 1}. \quad (8) \end{aligned}$$

The advantage of the second approach is that no parameters of the mutational model need to be estimated. All the power of the method is therefore concentrated on estimating parameters of the demographic/genealogical model. However, the conclusions obtained under this model may not be robust to deviations from the assumption of a uniform distribution. In particular, when the mutation rate is low, a bimodal distribution of allele frequencies is expected with most of the probability mass centered at zero and one, which is consistent with what is found in surveys of variation in nucleotide sequences in human populations (e.g., Hey 1997).

Still another possibility is to allow the gene frequencies to follow a beta distribution as expected under mutation-drift equilibrium in diallelic models. In that case, $\Pr_1(\{A, a\})$ would be given by:

$$\begin{aligned} \Pr_1(\{A, a\} | \alpha, \beta) &= \int_0^1 \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} f^{\alpha-1} (1-f)^{\beta-1} \binom{A+a}{A} f^A (1-f)^a df \\ &= \binom{A+a}{A} \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + A)\Gamma(\beta + a)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha + \beta + A + a)}, \end{aligned} \quad (9)$$

where α and β now are parameters of the model (this distribution is known as the beta-binomial distribution). A special case of this distribution is the symmetric beta distribution, $\alpha = \beta$. This distribution provides the equilibrium distribution under a neutral diallelic model with symmetric mutation rates when $\theta = 2\beta$, where $\theta = 4N_e\mu$ and μ is the mutation rate. The special case of $\beta = 1$ ($\theta = 2$) gives the uniform distribution. One of the advantages of this approach is that the beta distribution will fit most symmetric continuous $[0,1]$ density functions and hence may provide a good approximation when data from many loci are combined.

A final possibility is to use the observed allele frequencies as an estimate of the true allele frequencies. It is easy to show that the observed allele frequencies provide an unbiased estimate of the ancestral allele frequencies. However, as discussed in Nielsen et al. (1998), using the observed allele frequencies will bias the estimate toward lower divergence times and/or higher migration rates. Therefore, this approach cannot be recommended.

DISTINGUISHING MIGRATION FROM DIVERGENCE

The likelihood method presented in the previous section jointly estimates M , the rate of migration scaled by $2N$, and T , the time of population divergence, also scaled by $2N$. If a beta distribution is assumed as a prior distribution of allele frequencies before population divergence, the parameter of that distribution, β , is also estimated. In addition to parameter estimates, our method provides a framework for testing hypotheses about the history of the populations. There are two hypotheses that can be tested. The first is that $M = 0$, meaning that there has been no migration between the populations since they diverged. The second is that T is large, meaning that the historical association of populations is not the cause of their current genetic similarity. Both hypotheses can be tested using the likelihood ratio as a test statistic. The likelihood ratio is the ratio of the maximum likelihood when

both M and T are estimated to the maximum likelihood when one or the other parameter is constrained. The χ^2 approximation to the distribution of the likelihood-ratio test statistic is valid provided that a large number of loci have been sampled. For example, the hypothesis of $T = 5$ can be tested by comparing $-2 \log\{\max\{L_{T=5}(\beta, M)\}/\max\{L(\beta, M, T)\}\}$ to a χ^2 distribution with one degree of freedom. When the parameter being tested is at the boundary of the parameter space (e.g., $M = 0.0$), the likelihood-ratio test based on the usual χ^2 approximation will (asymptotically) be conservative. Under the null hypothesis, when the single parameter of interest is at the boundary of the parameter space, the likelihood-ratio test statistic is (asymptotically) distributed as a random variable that takes on a value from a χ_1^2 distribution with probability 0.5 and takes the value zero with probability 0.5 (Chernoff 1954; Feng and McCulloch 1992).

To evaluate the method for the data simulated under a uniform prior distribution of allele frequencies, we applied the method described above to jointly estimate T and M for each replicate. An example in which the true values are $M = 0.2$, $T = 1$ is shown in Figure 2. Here and later we assumed that a sample of 10 individuals ($n = 20$) was typed at 50 loci. Notice that reasonable estimates of M and T are obtained and that the hypothesis of $T = 0$ easily can be rejected provided we assume that $-2 \log\{\max\{L_{T=0}(\beta)\}/\max\{L(\beta, T)\}\}$ is distributed as discussed above. Notice that there appears to be some power to reject the hypothesis of $M = 0$. Another example for $M = 2.0$ is shown in Figure 3. In all simulations only variable loci are included, in accordance with the described statistical model.

We illustrate the application of this method in several other cases shown in Table 1. In each case, data were simulated under the true parameter values and then maximum-likelihood estimates were obtained with and without different constraints. It is impractical to carry out the analysis when T is constrained to very large values, but we found that the maximum likelihoods and estimates of M and β for larger values of T differed only slightly from the values obtained when $T = 5$. The reason is that, in the recursion described above, there is almost always only one remaining lineage at $T = 5$, so increasing T further has little effect on the results. Therefore, rejecting the hypothesis that $T = 5$ is, for practical purposes, equivalent to rejecting the hypothesis that T is $>$ than 5.

The results shown in Table 1 illustrate how the method is applied to different cases and suggest that reasonable results are obtained even with relative small sample sizes (50 unlinked loci from 10 individuals). Because of limitations of computer facilities, we were unable to run very large numbers of replicate simulations to carry out a complete power analysis, but we were able to run enough replicates to suggest that there is reasonable power, even for these small sample sizes. We concentrated on the last two cases in Table 1. Studies of variation in SNPs in humans indicate that most SNPs are in relatively low or high frequency (e.g., Hey 1997), which is consistent with the assumption of a small value of β , but we wanted to determine how important deviations from the assumed prior distribution are. We ran 30 sets of simulations for the four sets of assumptions corresponding to the last two cases in Table 1. For each parameter value, 10,000

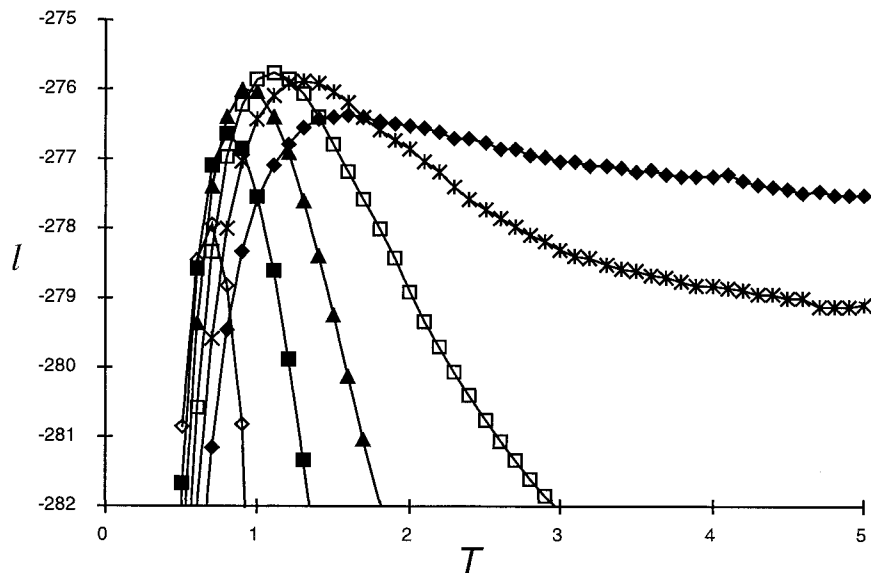


FIG. 2. The likelihood surface for a simulated dataset of 50 loci, each containing 20 genes. The dataset was simulated assuming a uniform distribution of ancestral frequencies, $M = 0.2$ and $T = 1$. (\diamond) $M = 0$; (\blacksquare) $M = 0.1$; (\blacktriangle) $M = 0.2$; (\square) $M = 0.3$; (\times) $M = 0.4$; (\blacklozenge) $M = 0.5$. For each parameter value 100,000 replicates were used to estimate the likelihood.

replicates were used to estimate the likelihood. In each case, the likelihood was optimized on a grid.

When the data were simulated using a beta distribution of allele frequencies before T , we found some power to reject the hypothesis that $M = 0$. If the true value of M was 0.5, then $M = 0$ was rejected in 18 of 30 replicates at a 95% level. If instead the true value of M was 2, $M = 0$ could be rejected in 22 of 30 replicates. The performance was worse when the simulated data did not fit the beta prior. When we assumed that 40 or 50 loci had a beta prior with $\beta = 0.1$ and 10 of 50 had a beta prior with $\beta = 1$, $M = 0$ could be rejected in only 12 of 30 replicates when $M = 0.5$ and in 13 of 30 replicates when $M = 2$.

APPLICATION TO DATA

To illustrate the method, we apply it to a previously published dataset of RFLP data. The dataset consist of 38 independent diallelic loci. It was previously described in Matullo et al. (1994) and Nielsen et al. (1998).

A likelihood surface was obtained for M , T , and β using 100,000 simulations for each datapoint. The likelihood for different values of β were evaluated in one run using importance sampling (e.g., Griffiths and Tavaré 1994a). The maximum-likelihood surface (maximized for β) for M and T is shown in Figure 4. Notice that there appears to be little power to distinguish different values of M . With the exception

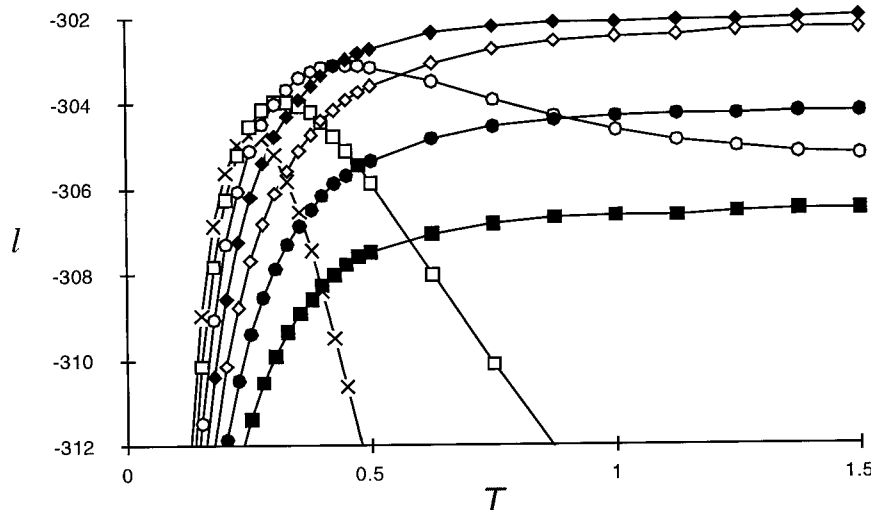


FIG. 3. The likelihood surface for a simulated dataset of 50 loci, each containing 20 genes. The dataset was simulated assuming a uniform distribution of ancestral frequencies, $M = 2$ and $T = 1$. (\times) $M = 0.0$; (\square) $M = 0.5$; (\circ) $M = 1.0$; (\blacklozenge) $M = 1.5$; (\diamond) $M = 2.0$; (\bullet) $M = 2.5$; (\blacksquare) $M = 3.0$.

TABLE 1. Estimates of parameter values for different simulated datasets. In all cases, the simulation program generated a sample dataset with 50 unlinked loci in 10 diploid individuals using the true values of the parameters, as shown. The carat indicates the maximum-likelihood estimates and L is the likelihood at the maximum-likelihood estimate of the unconstrained parameters. The constraints are the parameter values, if any, that are fixed in the determination of the maximum-likelihood estimates. For each parameter value 10,000 replicates were used to estimate the likelihood.

True values	Constraints	$M = 0.2$				$M = 2.0$			
		\hat{M}	\hat{T}	$\hat{\beta}$	$\ln(L)$	\hat{M}	\hat{T}	$\hat{\beta}$	$\ln(L)$
$T = 1.0, \beta = 1.0$	$\beta = 1$	0.15	1.2	—	-259.21	2.0	5.0	—	-295.04
	$M = 0, \beta = 1$	—	1.0	—	-260.54	—	0.2	—	-298.62
	$T = 5, \beta = 1$	0.35	—	—	-262.23	2.0	—	—	-295.04
$T = 1.0, \beta = 1.0$	none	0.15	1.2	1.0	-259.18	2.0	5.0	0.10	-295.04
	$M = 0$	—	1.0	4.0	-260.09	—	0.2	1.0	-297.66
	$T = 5$	0.35	—	∞	-262.23	2.0	—	0.10	-295.04
$\beta = 0.1, T = 1.0$	none	0.35	1.2	0.01	-274.20	2.0	5.0	0.01	-293.70
	$M = 0$	—	0.8	1.5	-277.33	—	0.2	0.8	-294.09
	$T = 5$	0.4	—	3.2	-277.55	2.0	—	0.01	-293.70
40 loci: $\beta = 0.1$	none	0.5	0.8	0.8	-273.19	3.0	5.0	3.2	-294.10
10 loci: $\beta = 1.0$	$M = 0$	—	0.8	1.6	-273.45	—	0.2	1.6	-296.75
$T = 1.0$	$T = 5$	0.4	—	0.01	-277.84	3.0	—	3.2	-294.10

that a lower bound for T can be established, very few conclusions regarding gene flow and divergence times can be reached. Apparently, there is not enough information in the data to distinguish effects of ongoing gene flow from population divergence. Another contributing factor may be inadequacies of the model. The simple models of migration and divergence traditionally used in population genetical inference cannot capture all the biological complexities of temporally varying population sizes and migration rates effecting real populations.

DISCUSSION AND CONCLUSION

The method we describe here provides a way to consider the combined effects of ongoing gene flow and historical association of populations on the frequencies of diallelic

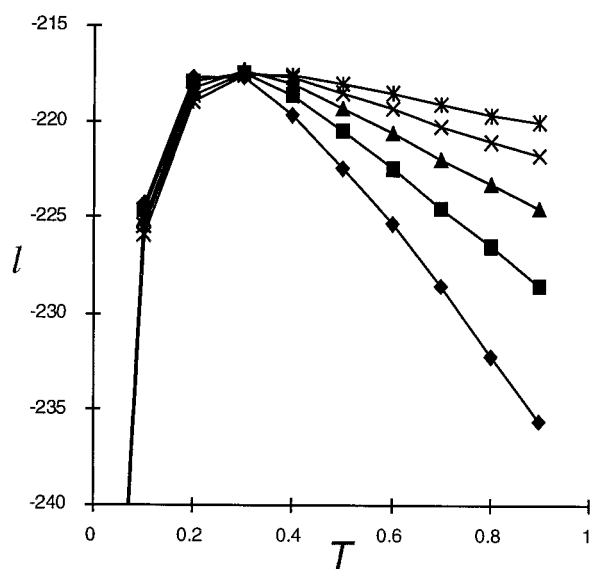


FIG. 4. The likelihood surface for a dataset of 38 human restriction fragment length polymorphism loci from an Italian and a Japanese population. (\diamond) $M = 0$; (\blacksquare) $M = 0.25$; (\blacktriangle) $M = 0.5$; (\times) $M = 0.75$; (\ast) $M = 1$.

marker alleles. This method, which is a generalization of the method described by Nielsen et al. (1998), is based on the general simulation approach introduced by Griffiths and Tavaré (1994a,b) but specialized to the case in which the mutation rate is assumed to be zero. The assumption of zero mutation rate is equivalent to assuming that mutation does not significantly affect marker frequencies during the time scale of interest, that is, since T , the time of population divergence.

There are three reasons for assuming no mutation. First, is that the mutation rates for SNP and RFLP markers are known to be very small. Second, by ignoring mutation, there are fewer parameters to estimate and fewer assumptions about the demographic history of the population. We do not need to estimate an average mutation rate or parameters describing the distribution of mutation rates among loci. Instead, we have to estimate a parameter (β) describing the distribution of marker frequencies before population divergence, but this has the advantage that this parameter could be estimated from the observed distribution of marker frequencies and that it does not require any assumption about the demographic history of the population before divergence. The third reason for ignoring mutation is the greatly increased computational efficiency. Including mutation would add a step to the recursions and would require simulation to times in the past before population divergence. Even when mutation is ignored, considerable simulation time was required to obtain accurate values of the likelihoods.

We show here that for diallelic models there is some power to distinguish models of strong migration and long divergence times from models of low migration and short divergence times without making unrealistic assumptions, but our preliminary power analysis suggests that relatively large numbers of loci will be needed to draw strong conclusions. The increasing use of SNPs and the increasing availability of efficient methods for screening populations will soon provide much larger sample sizes than are available today. Our results show that a formal likelihood method will be able to distinguish between ongoing gene flow and historical association in a way that cannot be done with simple summary

statistics such as F_{ST} . A much more extensive analysis of the power of this method under different conditions will be necessary to fully understand its applicability and limitations. Our goal is to show that such a method can be useful and that it deserves further study.

ACKNOWLEDGMENTS

This study is supported by National Institute of Health grant GM40282 to MS and in part by a fellowship to RN from the Danish Research Council and NSF grant DEB-9815367 to J. Wakeley.

LITERATURE CITED

- Bahlo, M., and R. C. Griffiths. 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* *In press*.
- Beerli, P. B., and J. Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152: 763–773.
- Chernoff, H. 1954. On the distribution of the likelihood ratio. *Ann. Math. Statist.* 25:573–578.
- Feng, Z., and C. E. McCulloch. 1992. Statistical inferences using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Stat. Prob. Lett.* 13:325–332.
- Griffiths, R. C., and S. Tavaré. 1994a. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46:131–159.
- . 1994b. Ancestral inference in population genetics. *Stat. Sci.* 9:307–319.
- Hey, J. 1997. Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* 14:166–172.
- Matullo, G., R. M. Griffo, J. L. Mountain, A. Piazza, and L. L. Cavalli-Sforza. 1994. RFLP analysis on a sample from northern Italy. *Gene Geo.* 8:25–34.
- Nath, H. B., and R. C. Griffiths. 1996. Estimation in an island model using simulation. *Theor. Popul. Biol.* 50:227–253.
- Nielsen, R. 1998. Maximum likelihood estimation of population divergence times and population phylogenesis under the infinite sites model. *Theor. Popul. Biol.* 53:143–151.
- Nielsen, R., J. Mountain, J. P. Huelsenbeck, and M. Slatkin. 1998. Estimation of population divergence times and population phylogenies in models without mutation. *Evolution* 52:669–677.
- Templeton, A. R., E. Routman, and C. A. Phillips. 1995. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* 140: 767–782.
- Wakeley, J. 1996. Distinguishing migration from isolation using the variance of pairwise differences. *Theor. Popul. Biol.* 49: 369–386.
- . 1998. Segregating sites in Wright's island model. *Theor. Popul. Biol.* 53:166–174.

Corresponding Editor: E. Zouros