A Cladistic Measure of Gene Flow Inferred from the Phylogenies of Alleles

Montgomery Slatkin and Wayne P. Maddison

Department of Zoology, University of California, Berkeley, California 94720

Manuscript received March 21, 1989
Accepted for publication July 14, 1989

ABSTRACT

A method for estimating the average level of gene flow among populations is introduced. The method provides an estimate of Nm, where N is the size of each local population in an island model and m is the migration rate. This method depends on knowing the phylogeny of the nonrecombining segments of DNA that are sampled. Given the phylogeny, the geographic location from which each sample is drawn is treated as multistate character with one state for each geographic location. A parsimony criterion applied to the evolution of this character on the phylogeny provides the minimum number of migration events consistent with the phylogeny. Extensive simulations show that the distribution of this minimum number is a simple function of Nm. Assuming the phylogeny is accurately estimated, this method provides an estimate of Nm that is as nearly as accurate as estimates obtained using F_{ST} and other statistics when Nm is moderate. Two examples of the use of this method with mitochondrial DNA data are presented.

THE extent of gene flow determines the extent to which different populations of a species are independent evolutionary units. There are a variety of indirect methods for estimating the average amount of gene flow from allozyme data (SLATKIN 1985). Methods based on Wright's F_{ST} and on private alleles provide robust estimates of Nm in a demic model, where N is the average deme size and m is the average migration rate, or Wright's neighborhood size in a continuum model (SLATKIN and BARTON 1989). However, these methods are not well suited for the analysis of DNA sequence data because they do not make full use of information in the data and because they require frequencies at several independent loci in order to provide reasonably accurate estimates. We will describe a new method that does use some of the additional information provided by DNA sequences. This method does not require complete sequences but does assume there is sufficient variation in restriction sites that an accurate phylogeny can be inferred for the segments of DNA sampled. We agree with AVISE *et al.* (1987) that the phylogenetic analysis of genetic diversity has the potential to offer considerable insight into processes governing population differentiation.

BACKGROUND

Data: Throughout, we will be concerned with the ancestry of a sample of nonrecombining segments of DNA from different individuals. Our method is designed to analyze data from studies of within-species variation in mitochondrial and chloroplast DNA such as those reviewed by AVISE *et al.* (1987). We do not yet know whether our method will be applicable to samples of recombining portions of nuclear DNA. For convenience, we will refer to a segment of nonrecombining DNA, such as the mitochondrial genome, as a gene, and assume that only one gene is sampled from each individual. If different segments of DNA were sampled from each individual, then our method would be applied separately to each.

We will assume that genes sampled have been examined for differences in DNA sequence either by direct sequencing or by using a battery of restriction enzymes. Given the differences in sequence, it is possible to reconstruct the phylogeny of the genes sampled by using one of the methods available. FELSENSTEIN (1988) discusses several alternative methods. It is likely that each gene would be found to be unique if enough of each sequence could be examined. In practice, however, genes sampled from different individuals sometimes appear to have the same haplotype. As we will discuss later, that does not pose a problem for our method but for now it will be more convenient to assume that each gene is unique and that a phylogeny can be reconstructed from the differences in sequence.

Coalescent processes: For selectively equivalent genes, KINGMAN (1982a, b) and others have shown that it is sufficient to consider only the direct ancestors of genes in a sample. This is known as the "coalescent" or "genealogical" approach in population genetics (TAVARÉ 1984). In following the ancestry of a sample of genes, either each gene in the sample is descended from a different gene in the previous generation, or

The publication costs of this article were partly defrayed by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

two or more genes are descended from one gene in the previous generation. In the second case, we will follow the convention and say that a "coalescent event" or "coalescence" has occurred. In the ancestry of a sample of genes, there will be a sequence of coalescent events until only one gene remains. Genes that are not direct ancestors of those in the sample do not need to be considered. Only the total population size and breeding system will affect the probabilities of coalescent events in each generation. The sequence of coalescent events will result in a treelike structure describing the ancestry of the genes sampled, making it natural to use phylogenetic methods.

Parsimony: The starting point for our analysis is a sample of genes drawn from two or more geographic locations. We will assume that a phylogeny of the genes has been inferred using differences in DNA sequence and not using any geographic information. We will assume that the samples are drawn from distinct geographic areas, with more than one individual sampled from each area, and will denote the sampling location by 1, 2, . . . , r. We will illustrate our method for $r = 2$. As shown in Figure 1, the sampling location can be regarded as an r state character that is associated with each gene sampled. We treat this character as a multistate unordered character on a tree and use a parsimony criterion to determine the minimum number of transitions, *i.e.*, migration events, consistent with the tree.

The algorithm for the case with $r = 2$ is illustrated in Figure 1. We assign to each terminal node the set {1} if the corresponding gene is from location 1, and {2} if from location 2. The procedure is then to move down the tree toward the root recursively, assigning sets of states to internal nodes. The possible state sets of internal nodes are {1}, {2} and {1, 2}. At each step in the recursion, the rule for joining two sets is a straightforward majority-rules voting procedure. If possible, the ancestor's state set is made of states that occur in both state sets joined. For example, a {1} joining a {1, 2} implies a state set of {1} for that node. If a {1} joins a {2}, the state set of the node is {1, 2}. The state sets so assigned to the nodes do not necessarily indicate the parsimony reconstruction of ancestral state. Additional calculations are needed to readjust these state sets to achieve a final parsimony reconstruction (FARRIS 1970). Nonetheless, this algorithm does allow us to determine the minimum number of state changes, as noted by FITCH (1971). Among the six possible joinings that can occur with $r = 2$, only one of them, a {1} joining a {2} requires that we assume that a state change, *i.e.*, a migration event, has occurred. By summing these joinings we will obtain the minimum number of migration events, which we will denote by s, consistent with the data.

This algorithm can be easily generalized to more

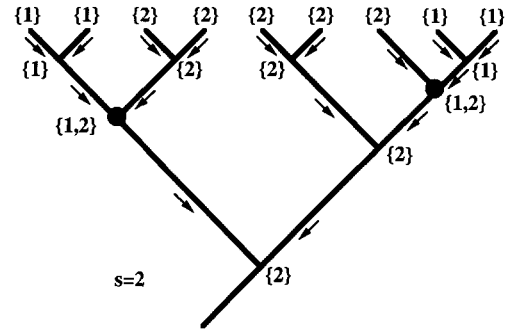

FIGURE 1.— An illustration of how the algorithm of FITCH (1971) is used to compute s from a phylogeny of nine genes. Four genes are assumed to be sampled from location 1 and five genes from location 2. The numbers in braces indicate the inferred state of the nodes as the calculation is being performed. Two numbers in braces indicate that the state of the node is initially ambiguous. That ambiguity may be resolved by the states of earlier nodes. In this example both ambiguous nodes would be assigned to state 2 under the parsimony assumption. For our purposes, however, the resolution of ambiguous states is unneeded because the value of s is not affected.

than two sampling locations. With three locations, 1, 2 and 3, there are seven kinds of state sets that can be assigned {1}, {2}, {3}, {1, 2}, {1, 3}, {2, 3}, and {1, 2, 3}. With r states, there are $2^r - 1$ possible state sets for each node, namely all the nonempty subsets of {1, 2, . . . , r}. The rules for assigning state sets to nodes are the natural generalizations of the rules used for $r = 2$. The ancestor's state set is made up of any states that occur in both state sets joined; if no states are present in both then a migration event must have occurred and the ancestor's state set is made up of all the states in both state sets. This algorithm can be rephrased in terms of set operations: when two genes join, the state set of the ancestor is assigned to be the intersection of the state sets of the two genes unless that intersection is empty, in which case there is assumed to be a migration event (*i.e.*, s is increased by 1) and the state set of the ancestor is assigned to be the union of the two state sets.

This algorithm assumes in effect that every deme is equally accessible from every other deme, which is equivalent to assuming an island model of migration. It is possible to consider other models of migration that assume that some migration events are more difficult than others. For example, a linear array of three populations would require the assumption that migration directly between the end populations would count as two migration events instead of one. There is a generalization of the algorithm in Figure 1 for a general matrix of transitions, but we will not consider that problem here.

Our method is then very simple. The phylogeny of genes combined with the geographic locations indicates the minimum number of migration events, s, in the history of those genes necessary for their current geographic distribution to be consistent with their

phylogeny. The problem is to use s to estimate Nm or some other combination of parameters describing the amount of gene flow affecting the species from which the samples were taken. Our overall approach is similar to one proposed by HUDSON and KAPLAN (1985) for using the minimum number of recombination events to estimate the product of population size and recombination rate.

Simulation program: To determine the distribution of s, the minimum number of migration events, under different assumptions about gene flow and population structure, we developed a Monte Carlo simulation program. We assumed that the genes sampled were selectively neutral, which allowed us to use the coalescent approach described above. TAKAHATA (1988), SLATKIN (1989), and TAKAHATA and SLATKIN (1989) have already used this approach for examining the consequences of gene flow on samples of genes.

Our simulation program assumed a collection of d populations each of size N. We assumed that n_i genes were sampled from deme i ($i = 1..r$, $r \leq d$). In each generation, there were two steps, migration and the production of gametes. We assume that the immigration rate, m, is the same in each population, so the probability that a gene is an immigrant is m and the probability that it is not is $1 - m$. If there are n_i genes in deme i whose ancestry we are concerned with and migration is assumed to occur at the gamete stage, the probability that j of them are immigrants is given by a binomial distribution with parameters m and n_i:

$$\Pr(j \text{ immigrants}) = \binom{n_i}{j} m^j (1 - m)^{n_i - j} . \qquad (1)$$

If m is sufficiently small, $\Pr(j = 0) \approx 1 - n_i m$, $\Pr(j = 1) \approx n_i m$ and $\Pr(j > 1) = O((n_i m)^2)$, where $O(.)$ indicates the order of magnitude of the terms. When a migration event occurred we assumed that the immigrant had a probability of $1/(d - 1)$ of coming from each other deme.

At the reproduction stage, we modeled a haploid population of N adults and assumed that gametes were sampled with replacement to form the next generation. We assumed that $N >> n$, the number of genes, at all times, in which case the probability that n genes are descended from n different parents (*i.e.*, there was no coalescent event) is approximately $1 - n(n - 1)/(2N)$, and the probability that they are descended from $n - 1$ parents (*i.e.*, there was a coalescent event) is approximately $n(n - 1)/(2N)$ (KINGMAN 1982a). When $N >> n$, the probability of two or more coalescent events is of order $(n/N)^2$. TAKAHATA (1988) calls this approximation for the coalescent process the "diffusion limit."

In each replicate simulations, we specified the parameters of the population: N, the population size of each deme; m, the immigration rate for each deme; and d, the number of demes. We assumed that n_i genes were sampled from deme i ($i = 1..r$) and labeled each gene with the deme from which it was sampled. Then we simulated the history of the sample of genes as affected by migration and coalecent events for t generations in the past. After t generations, we assumed that the d demes were descended from a single panmictic deme. The size of the ancestral deme does not matter because we are not concerned with the times of occurrence of coalescent events, only with which genes coalesced. We continued to simulate the sequence of coalescent events in the panmictic population until only one gene was left.

During each replicate, we counted the number of coalescent events between genes whose identities indicated that they were sampled from different populations. That is, we used the algorithm described in the previous section to compute s, the minimum number of migration events necessary to make the ancestry of the genes sampled consistent with their current geographic distribution. Each replicate simulation yielded a single value of s, and a set of replicates provided an estimate of $p(s)$, the distribution of s for a given parameter values and sample sizes. We will be particularly concerned with the mean and variance of s. Throughout, we will denote the average of s in a set of replicate simulations by $\bar{s}$, and the average squared deviation from $\bar{s}$ by $\hat{\sigma}_s^2$. The quantities $\bar{s}$ and $\hat{\sigma}_s^2$ are estimators of the mean and variance of $p(s)$.

We used two slightly different simulation programs. In the first, we allowed for the possibility of more than one migration event per generation, which let us assume relatively large values of m. In the second program, we assumed that m was sufficiently small that no more than one gene per generation could be an immigrant. The first program was quite slow but we used it to demonstrate that the simulation results for high migration rates were indistinguishable from exact analytic results for samples of the same size from a single panmictic population. The results from the two programs were in complete agreement when the same parameter values were used. The results described below were all obtained using the second program.

Two demes in a population at equilibrium ($d = 2$, $r = 2$): We found that the distribution of s, $p(s)$ depends only on the product Nm as long as N is much larger than the number of genes sampled. Figure 2 shows some results for a population with only two demes which have been separated for long enough that the results do not depend on t, the time of separation. For $Nm > 1$, $p(s)$ is approximately normal and the variance is approximately constant, as shown in Figure 3, so $\bar{s}$ and $\hat{\sigma}_s^2$ are useful descriptors of $p(s)$. Figure 4 shows a graph of $\bar{s}$ *vs.* Nm.

We can use these results to estimate Nm from s

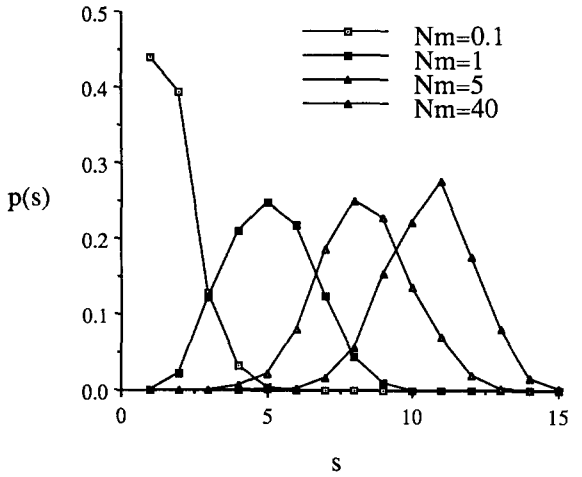

FIGURE 2.—Comparison of $p(s)$ for different values of Nm. In all cases, 16 alleles were sampled from each of two demes ($r = 2$) and only two demes of size $N = 10{,}000$ were assumed to be in the population ($d = 2$). The curves are based on 1000 replicate simulations each and the population were assumed to have separated $t/N = 50$ generations in the past.

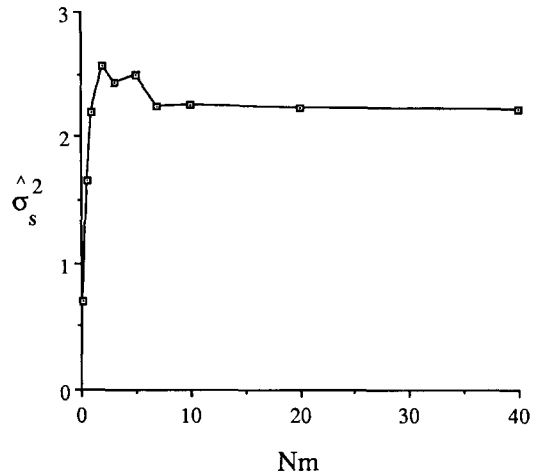

FIGURE 3.—The variance of s, $\hat{\sigma}_s^2$ as a function of Nm. The parameter values for the simulations were the same as in Figure 2 (*i.e.*, $r = 2$, $d = 2$ and $N = 10{,}000$).

computed for a particular sample of genes. The curve in Figure 4 appropriate for the sample sizes provides an estimate of Nm and indicates the degree of confidence in the resulting estimate. For this method to be useful we will have to determine how the resulting estimate of Nm depends on the number of demes in the population and the sample sizes.

More than two demes in population ($d > 2$, $r = 2$): If the numbers of genes sampled are fixed, it is relatively easy to guess the results when there are more than two demes in the population. With only two demes, an emigrant from one deme has to go to the other deme, after which it is as liable to coalesce in that deme as any of the genes that were not immigrants. If there are more than two demes, an emigrant from one deme has a probability $1/(d-1)$ of arriving in the other deme from which samples are taken and a probability of $(d-2)/(d-1)$ of arriving in one of

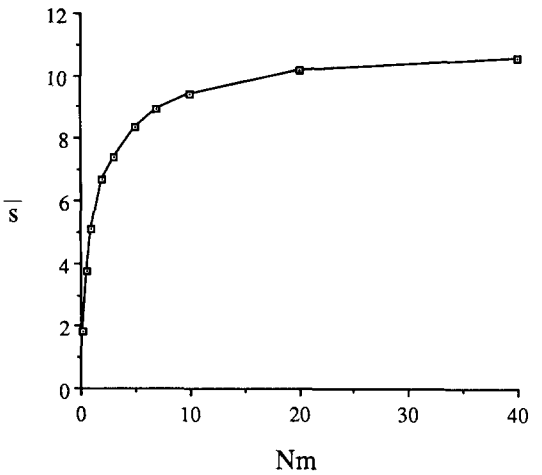

FIGURE 4.—The average of s, $\bar{s}$, as a function of Nm. The parameter values for the simulations were the same as in Figure 2 (*i.e.*, $r = 2$, $d = 2$ and $N = 10{,}000$).

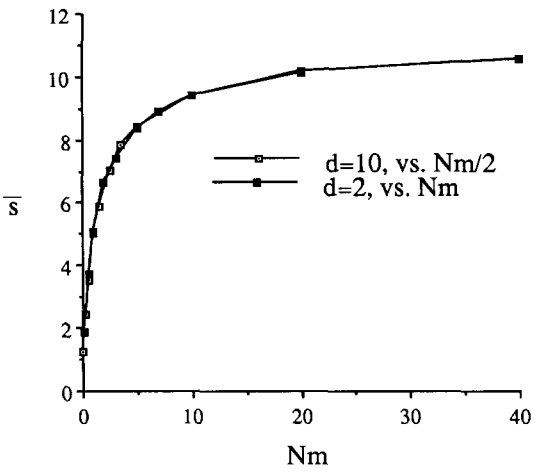

FIGURE 5.—$\bar{s}$ plotted against Nm for $d = 2$ and $\bar{s}$ plotted against $Nm/2$ for $d = 10$. Otherwise, the parameter values are the same as in Figures 2–4 ($r = 2$ and $N = 10{,}000$).

the demes from which were no samples. In the latter case, the emigrant will either have to migrate again to one of the two demes sampled before it coalesces or coalesce in one of the $d - 2$ other demes. In the island model, the gene we are focusing on is equally likely to return to the deme it started from or to the other deme sampled. Similarly, if the coalescence occurs in another deme, the other emigrant is equally likely to have come from either of the demes sampled. Hence, emigration to the other deme sampled will be equivalent to emigration in the two deme case but emigration to another deme will be only half as effective as emigration in the two deme case. Therefore, with d demes, a migration rate of m would be equivalent to a migration rate of $m/(d-1) + m(d-2)/[2(d-1)]$ or $md/[2(d-1)]$ between the two sampled demes. Figure 5 confirms this intuitive argument by showing that s plotted against Nm for $d = 2$ is nearly identical to s plotted against $Nm/2$ for $d = 10$. Hence if the island model is an appropriate description of the population structure and if d is assumed to be

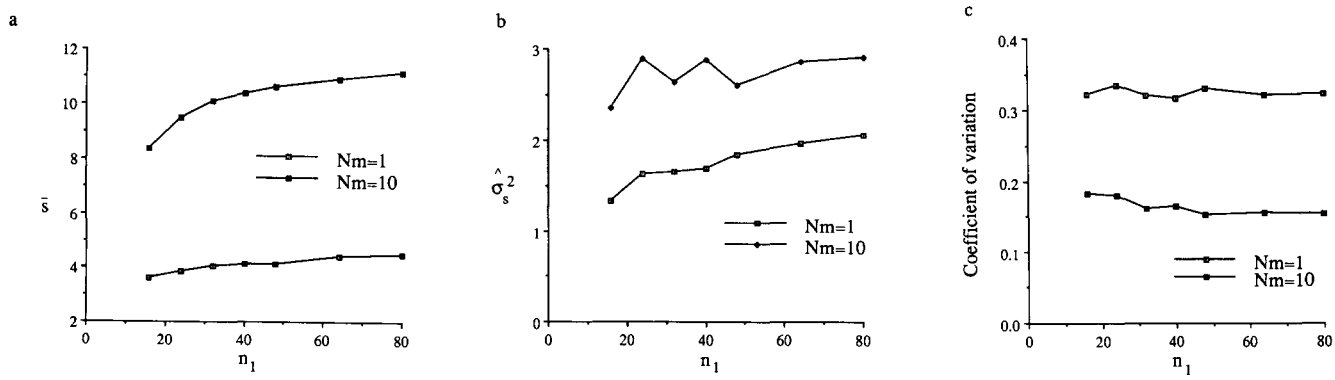

FIGURE 6.— a, $\bar{s}$ plotted against n_1, the number of alleles sampled from one of two demes, given that $n_2 = 16$. In all cases, $N = 10{,}000$ and 1,000 replicates were run. b, $\hat{\sigma}_s^2$ plotted against n_1 from the same simulations as in part a. c, $\hat{\sigma}_s/\bar{s}$ plotted against n_1.

reasonably large ($d \geq 5r$), then the estimate of Nm obtained with our method will not depend significantly on d.

In applications, there are probably many more demes in a population than are sampled, so the dependence on d is unimportant. We will assume that d is enough larger than r that its value can be ignored. In general, results for simulations with $d > 5r$ are indistinguishable from those for $d = 5r$.

Dependence on sample sizes ($r = 2$): If sample sizes in both demes are increased by the same multiple, both $\bar{s}$ and $\hat{\sigma}_s^2$ increase as linear functions of sample size. Hence the coefficient of variation of $p(s)$ decreases with the square root of sample size, as intuition would suggest.

If different numbers of genes are sampled from two different populations, then the results depend most strongly on the smaller of the two sample sizes. Some typical results are shown in Figure 6a. Even though the number of genes sampled from one of the locations increases by a factor of 5, $\bar{s}$ increases by only about 20%. Figure 6b shows that the variance exhibits the same weak dependence on sample size and Figure 6c shows that $\bar{s}/\hat{\sigma}_s$ is apparently independent of the larger sample size. Therefore, when samples are taken from only two locations, the accuracy of the estimate of Nm depends primarily on the smaller of the two sample sizes.

More than two populations sampled ($d \geq r > 2$): When more than two populations are sampled, the analysis and the interpretation are more complicated but the results are relatively simple. Although finding the minimum number of migration events can be done by hand for an arbitrary number of locations sampled, it is probably easier to use a computer program when more than two states are present. The program MacClade (written and distributed by W. P. MADDISON and D. R. MADDISON) has a subroutine that carries out this calculation.

If there are equal sample sizes from each location, then $\bar{s}$ increases approximately linearly with r, the number of locations sampled, as shown in Figure 7. We can understand why this is so by first considering the case with three demes sampled. If we focus on the samples from location 1 and group the samples from location 2 and 3 into a single sample, 2′, then we can predict the resulting value of $\bar{s}$ using the two-population results given above. For example, if 16 genes are sampled from each of three locations, we would use the data on which Figure 6a is based to predict $\bar{s}$ for the (1, 2′) grouping to be 4.022 if $Nm = 1.0$. Let this value of $\bar{s}$ be denoted as $\bar{s}^*$. The fact that the 2′ genes are a mixture of the genes from locations 2 and 3 makes no difference if we are concerned with the number of coalescent events between genes from location 1 and genes from the other locations. We can then consider the coalescent events between genes from locations 2 and 3 and ignore the fact that other coalescent events are occurring with genes from location 1. This would give us a second value of $\bar{s}$, which we will denote by $\bar{s}^{**}$. If 16 genes are sampled from each of two demes, then $\bar{s}^{**} = 3.581$. The value of $\bar{s}$ is not exactly $\bar{s}^* + \bar{s}^{**}$ because there are some trees for which this partitioning of the tips leads to an incorrect final value of s.

By extrapolation, we can see why $\bar{s}$ is approximately a linear function of r. To predict $\bar{s}(r)$ as a function of r, first consider one location separately and write $\bar{s}(r) = \bar{s}^* + \bar{s}^{**}$, as in the case with $r = 3$, where $\bar{s}^*$ is the value for $r = 2$, with n genes sampled from one and $(r - 1)n$ genes from the other. Then we note that $\bar{s}^{**} \approx \bar{s}(r - 1)$ and recall that in the previous section $\bar{s}$ is almost independent of the larger of the two sample sizes when samples are taken from only two locations (*cf.* Figure 6). Hence $\bar{s}^*$ is nearly independent of r which implies $\bar{s}(r)$ is nearly linear in r, as we have found.

When different numbers of genes are drawn from more than two locations, we can use the same line of argument to predict $\bar{s}$. The results are more complicated but similar in character to the previous results. If the sample sizes are almost equal, the results are

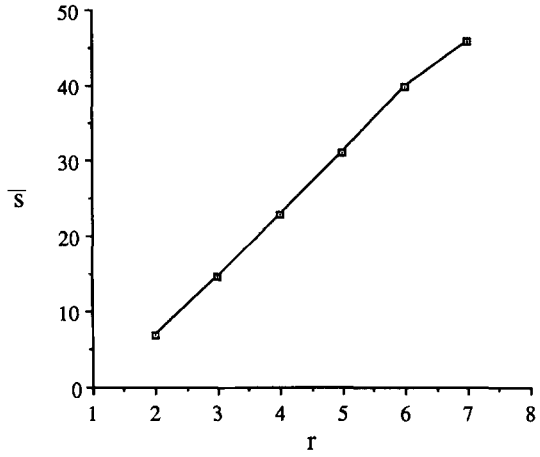

FIGURE 7.—$\bar{s}$ plotted against r, the number of demes sampled, with n_i, the number of alleles sampled from each deme held fixed. In all cases, $n_i = 16$ ($i = 1, \ldots, r$), $N = 10{,}000$, $d = 20$, $Nm = 5.0$ and $t/N = 50.0$.

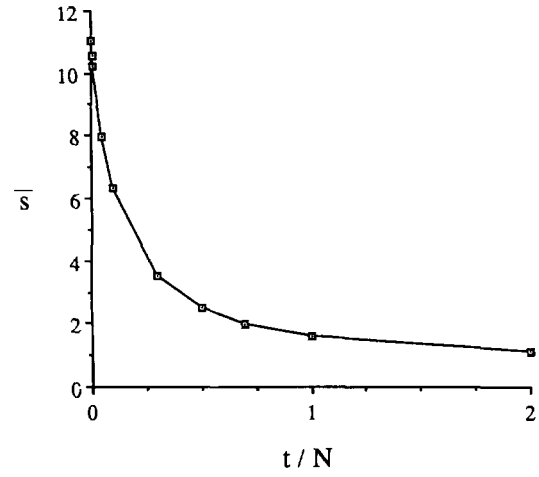

FIGURE 8.—$\bar{s}$ *vs.* t/N for the radiation model as described in the test. In all cases, $N = 10{,}000$, 16 alleles were sampled from each of two demes ($r = 2$) and the alleles were in a single panmictic deme until t/N in the past, after which there was no migration among the demes. In the radiation model, the number of demes after the cessation of migration is unimportant.

expected to be nearly the same as in the case with equal sample sizes. If the sample sizes are quite different, then the greatest contribution to the overall value of $\bar{s}$ will come from the largest sample sizes, although the smaller sample sizes will affect the variance.

Populations not at an equilibrium: So far we have assumed that the samples have been drawn from a collection of populations that have been separated for long enough that an equilibrium has been achieved between gene flow and genetic drift. In that case, common ancestry of genes in different populations is due to past migration. If populations have been derived from an ancestral population in the recent past, then common ancestry may not reflect ongoing gene flow but instead the historical association of the populations sampled.

To consider the effects of historical association, we considered the extreme case in which a single panmictic population gives rise to several independent populations at a time t in the past. The size of the original population does not matter because, before t, genes coalesce independently of the locations from which they were sampled. The size of that population would not affect the topology of the resulting phylogeny, only the branch lengths, which we are not considering here. For simplicity, we will assume that the sizes of the derived populations are all the same, N. We found that $p(s)$ and hence $\bar{s}$ depends on only the ratio t/N. Figure 8 shows some of our results.

There is in effect an equivalence between the divergence time, t/N in this model, and Nm in the equilibrium model with gene flow. Figure 9 illustrates this equivalence. These results indicate that in principle, there is no way for our method to distinguish between phylogenies that result from gene flow and from historical association. Whether other methods such as those that use branch lengths in additional to topology can do so is not yet known.

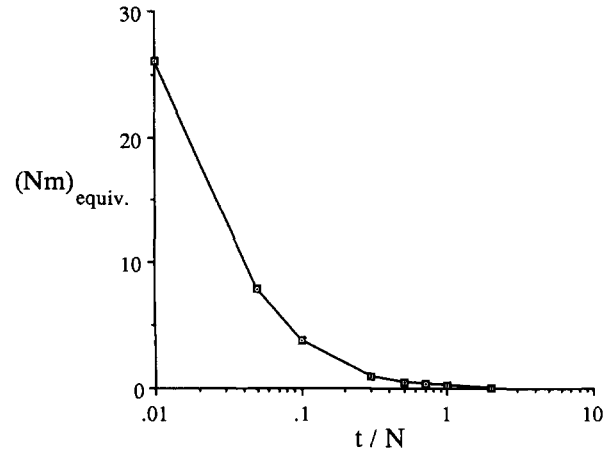

FIGURE 9.—The value of Nm in a population at equilibrium that yields the same value of s as in the simulations of the radiation model. The values plotted were obtained by using the values of $\bar{s}$ from the simulations shown in Figure 8 and using those values to estimate Nm for the same sample sizes.

ESTIMATING Nm FROM s

The above results suggest that an estimate of Nm can be obtained if s, the minimum number of migration events, can be computed from a phylogeny of genes sampled. Such an estimate of Nm depends on a number of assumptions, the most important of which are that the genes sampled are selectively equivalent and that the underlying population structure is well approximated by an island model. Obtaining a good estimate of Nm using our method is not easy because we have not been able to obtain an analytic expression for $\bar{s}$ as a function of Nm and the sample sizes. Instead, an estimate has to be based on our simulation results.

We have written a program to estimate Nm for arbitrary sample sizes and we will describe it below, but useful approximate estimates can easily be ob-

TABLE 1

Values of $\bar{s}$ and $\hat{\sigma}_s$ obtained from simulations described in text

	n					
	8		16		32	
Nm	$\bar{s}$	$\hat{\sigma}_s$	$\bar{s}$	$\hat{\sigma}_s$	$\bar{s}$	$\hat{\sigma}_s$
0.1	1.17	0.39	1.24	0.48	1.32	0.53
0.5	2.00	0.79	2.45	0.99	3.00	1.14
1.0	2.74	0.96	3.56	1.18	4.53	1.44
2.0	3.47	1.06	5.02	1.35	6.72	1.66
3.0	3.82	1.05	5.86	1.43	8.45	1.86
5.0	4.39	1.07	7.03	1.47	10.63	1.91
7.0	4.58	1.06	7.82	1.58	12.18	2.11
10.0	4.95	1.05	8.40	1.53	13.71	1.95
20.0	5.18	1.10	9.42	1.51	16.61	2.18
40.0	5.43	1.03	10.12	1.47	18.56	2.18

In each case, n alleles were drawn from each of two demes in an island model containing a total of ten demes. In each case, N = 10,000, $t = 50N$, and 1000 replicates were run.

tained. For samples from only two locations, a rough estimate of Nm can be obtained by interpolation from our simulation results. Table 1 presents some relevant values. Because $\bar{s}$ is so weakly dependent on the larger of the two sample sizes, little information is lost by equalizing the sample sizes and ignoring some of the genes in the larger of the two samples. To illustrate the use of Table 1, assume that 28 genes were sampled from one location and 35 from the other. The first step would be to reconstruct the phylogeny of all 63 genes and then remove 7 genes randomly from the larger sample. (Note that it is possible to get a somewhat different result if 7 genes are removed first and then the phylogeny is reconstructed. We feel it is preferable to reconstruct the phylogeny using all available information, and then to remove genes, although little or no difference would be expected from the two procedures.) Then the value of s is computed for the phylogeny. If the method of phylogeny reconstruction yields several equivalent phylogenies, s should be computed for each. Then the estimate of Nm can be obtained from Table 1 by interpolation. For example, if $s = 7$, then the estimate of Nm is approximately 3.9.

For samples from more than two locations, the first step would be to equalize the sizes of the largest samples, possibly by discarding some small samples. For example with sample sizes of 23, 34, 25, 6, 29, the phylogeny of all the genes would first be inferred. Then the data could be reduced to four samples of size 23 with the sample of size 6 ignored. An equivalent value of s for two sampling locations is obtained by multiplying this value of s by $2/r$, in this case $1/2$. The resulting value is then used with the values in Table 1 to obtain an estimate of Nm. For example, if the value of s from the samples from four locations is 19, then a rough estimate of Nm is 8.7.

The procedure based on the values in Table 1 is only approximate and usually requires that small sample sizes be discarded. More accurate estimates of Nm using all samples can be obtained by carrying out a series of simulations tailored to the particular sample sizes. We have written a program that will carry out the necessary simulations and produce an estimate of Nm given s and the sample sizes. The program assumes that d, the number of demes in the island model, is $5r$, the number of locations sampled. This program runs on a Unix system and we will distribute it upon request. It is very time consuming to run for each value of s, so we do not recommend its use unless free computational facilities are available. This program is written in Pascal and should be adaptable to other computer systems with some minor changes.

Confidence limits on estimates of *Nm*: Because our simulation results estimate $p(s)$ we can use this distribution infer confidence limits on the estimate of Nm. For $Nm > 1$ and roughly equal sample sizes, the distribution is nearly normal so $\hat{\sigma}_s$ usually adequate. Given the sample sizes and s, Nm is first estimated using Table 1. Table 1 also provides values of $\hat{\sigma}_s$ for samples of equal size from two populations. Then Nm can be estimated for the observed value of s plus or minus $2\hat{\sigma}_s$ (if 95% confidence limits are desired). In our simulation program, 95% confidence limits are determined using the simulated distributions of s, rather than assuming normality.

For example, assume that 16 genes are sampled from each of two locations and that $s = 7$. The resulting estimate of Nm is approximately 5 and the value of $\hat{\sigma}_s$ is 1.47 for this value of Nm and these sample sizes. The estimates of Nm for $s = 4.06$ ($= 7 - 2\hat{\sigma}_s$) is 1.3 and for $s = 9.94$ ($= 7 + 2\hat{\sigma}_s$) the estimate is 34.9. Therefore the confidence interval for the estimate of Nm is (1.3, 34.9). It is not symmetric about the estimate of Nm because Nm is a nonlinear function of $\bar{s}$.

One question is, if the total number of genes sampled that can be examined fixed, how many geographic locations should be sampled? To consider a specific example, assume that the total number of genes that can be examined is 32. If 16 genes are sampled from each of two locations then the confidence interval obtained using the complete distribution is (0.3, 8.0) when the estimate of Nm is 2.0 ($s = 5$). If 8 individuals are sampled from each of 4 locations and $s = 11$, the estimate of Nm is 2.1 and the confidence interval is (0.7, 6.8), and if 4 individuals are sampled from each of 8 locations and $s = 18$, then the estimate of Nm is 2.2 and the confidence interval is (0.8, 13.0). These results suggest that there is a slight preference to sampling more individuals from fewer locations but the effect is not strong enough to warrant distorting a sampling scheme that is devised primarily for other purposes.

We can compare confidence intervals obtained using this method with confidence intervals for indirect methods applied to allozyme data. Slatkin and Barton (1989) used F_{ST} and other indirect methods on simulated data. They considered samples from 10 polymorphic loci in 25 individuals from 10 demes, which represents typical published data sets. If G_{ST} = 0.136, the resulting estimate of Nm is 1.59 and the 95% confidence interval is (0.48, 2.70) (assuming a normal distribution of the estimates) (Slatkin and Barton, Table 1). If instead G_{ST} = 0.022, the estimate of Nm is 11.0 and the 95% confidence interval is (11.0, 12.9). To compare these confidence limits with those in the present method, assume 32 genes are drawn from each of 2 demes. Using our simulation program for s = 5, the estimate of Nm is 1.2 and the 95% confidence interval is (0.3, 2.8) and if s = 14, the estimate of Nm is 10.7 and the confidence interval is (4.2, 48.3).

For lower levels of gene flow the confidence intervals are nearly the same using the two methods even though with our method, we assumed that only two locations were sampled, but for higher levels of gene flow the confidence interval for the present method is substantially larger. Larger numbers of locations sampled would reduce the confidence interval somewhat. If 16 genes are sampled from each of 10 demes and s = 45, then our similation program indicates that the estimate of Nm is 1.6 and the 95% confidence interval is (1.1, 2.0).

For higher levels of gene flow, the confidence interval using this method is still larger than that obtained using F_{ST}, even with 10 demes sampled. For example, if 16 genes are sampled from each of 10 demes and s = 93, then our simulation program indicates that the estimate of Nm is 12.1 and the 95% confidence interval is (7.8, 21.1). Surprisingly, the confidence interval is not much smaller than that for a comparable case with only two locations sampled. The reason that the size of the confidence interval increases with Nm is that the graph of $\bar{s}$ *vs.* Nm flattens out for larger values of Nm so slight changes in $\bar{s}$ leads to large changes in the estimate of Nm.

EXAMPLES

We illustrate our method by applying it to data on two species of freshwater fish, *Lepomis punctatus* and *Lepomis gulosus*, studied by BERMINGHAM and AVISE (1986). These two species are found in streams along the Atlantic and Gulf of Mexico coasts of the southeastern United States. BERMINGHAM and AVISE used 14 to 17 endonucleases to assay the mtDNAs from each of the individual collected from several different rivers in their range. They constructed phylogenies of the mtDNAs for each of the species using both parsimony and phenetic methods. In both species and in two others they studied, they found that the phylogenies of the genes (clones) were partly concordant with the locations of the samples. In particular they found that samples taken from South Carolina, Georgia and most of Florida form one clade and samples from the western panhandle of Florida west to Louisiana to form a second clade. They concluded that there was no gene flow between those two regions and, because approximately the same boundary was found in all four species they studied, that this boundary represents a former biogeographic barrier to these species. SLATKIN (1989) used these data to place an upper bound on the extent of gene flow between the two regions. For one of the other species examined, *Amia calva*, $Nm < 0.2$ with 95% confidence. Similar values were found for the other three species including the two discussed here.

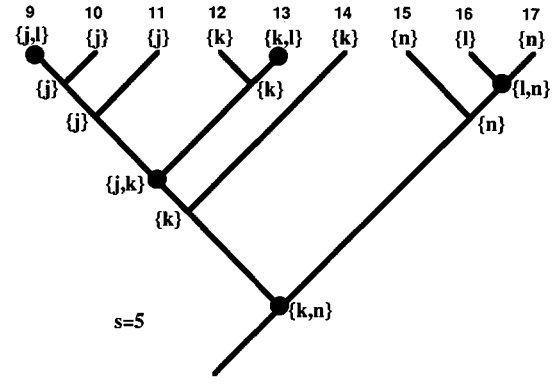

FIGURE 10.—Part of the cladogram of mtDNAs from the western samples of *L. punctatus* studied by BERMINGHAM and AVISE (1986, Figure 5). The numbers at the tips are the clone numbers and the letters indicate the river drainages in which each clone was found. Our notation follows that of BERMINGHAM and AVISE (1986). The solid circles indicate where migration events were counted.

We can use BERMINGHAM and AVISE's (1986) data to estimate the amount of gene flow among different river drainages within each region. We considered the western samples only and combined samples from different locations in the same drainage. The sample sizes for *L. punctatus* were 9 (*j*), 9 (*k*), 7 (*l*) and 5 (*n*), where the letters in parentheses indicate Bermingham and Avise's code for each river drainage. Both the phenogram constructed using a distance method and a cladogram constructed using Wagner parsimony indicated that s = 5 (BERMINGHAM and AVISE, 1986, Figure 5).

Figure 10 shows how we calculated s from their cladogram. One feature of the calculation is worth noting because it commonly occurs in such data. Given the resolving power of their method, BERMINGHAM and AVISE (1985) distinguished 17 distinct mtDNA "clones." Whether those clones could be further subdivided using more restriction enzymes or complete DNA sequences is currently unknown. As shown in Figure 10, two of these clones, 9 and 13, were present

in two of the drainages. We count one migration event for each of these clones at the tips of the tree, as indicated. We are treating the same clone found in two different locations as being different but forming a monophyletic group. According to our method of computing s, one migration event must have occurred after the two clones diverged. With this assumption, it does not matter how different in sequence members of the same clone actually are as long as they are still each other's closest relative.

With these sample sizes and $s = 5$, our simulation program estimated Nm to be 0.2. Using our simulation program we found that the upper confidence limit is 0.7 and the lower limit is 0. The lower limit cannot be taken literally because if Nm were actually 0, then $p(s)$ would be a spike at $s = 1$. We used 1000 replicates and the program took approximately 2 hr on a Sun 3/75 workstation. Larger values of s would imply higher values of Nm and would run much more slowly. If this low value of Nm represents the historical association of the populations in the different drainages rather than ongoing gene flow, then Figure 9 indicates that $t/N \approx 1$.

In *L. gulosus*, the results are similar. The sample sizes for the western samples are 6 (*j*), 6 (*k*), 12 (*l*), 4 (*m*) and 2 (*n*). Both the phenogram and the cladogram (BERMINGHAM and AVISE, 1986, their Figure 11) indicate that $s = 6$. This also leads to an estimate of Nm of 0.2 and an upper confidence limit of 0.8 and a lower limit of 0. The similarity of these estimates with those for *L. punctatus* could indicate a similar history of the two species but that conclusion would assume that the effective population sizes in each of the drainages were approximately the same.

DISCUSSION

Accuracy of the phylogeny: We have assumed that the phylogeny is accurately estimated from the DNA sequences or restriction site polymorphisms available. The accuracy of different methods for phylogeny reconstruction depends on whether particular assumptions about evolutionary processes are met (FELSENSTEIN 1988). Errors in the reconstruction of the phylogeny of the genes sampled would lead to inaccurate values of s. We think that most types of errors would tend to overestimate s and hence lead to overestimates of Nm, because most types of errors would tend to randomize the inferred phylogeny. We recommend that two or more different methods for phylogeny reconstruction be used. If estimates of Nm using different phylogenies are similar, then there is reason to have confidence in the average estimate. If estimates of Nm are quite different, then additional effort will be required to resolve the phylogeny before our method can be used with confidence. Similarly, if a particular method for inferring the phylogeny yields several phylogenies that fit the data equally well, then our method should be applied to each.

FELSENSTEIN (1985) has suggested a method for providing confidence limits on phylogenies. The result from applying his method is a phylogeny with multifurcations which represent possible bifurcations that cannot be distinguished. If our method is applied a tree with multifurcations, the resulting value of might be too small because the parsimony criterion we are using will assume the minimum number of migration events at each multifurcation (MADDISON 1989). If the actual phylogeny required larger values of s, then the resulting estimate of Nm would be too small.

A similar problem arises if genes sampled from different individuals appear not to be unique. For example, it is common in studies of variation in mtDNA to be unable to distinguish some individuals. More than one individual is often recorded as having the same haplotype, as in the BERMINGHAM and AVISE (1986) data. For our method to be applied, it must be assumed that individuals with the same haplotype form a single clade. If more than two individuals have the same haplotype, then there is in effect a mulitfurcation in the phylogeny. As in the case of other multifurcations, we recommend the resolution leading to the smallest value of s, in keeping with the parsimony criterion. If many individuals with the same haplotype are found, however, and particularly if they are found in several different locations, then our method will probably not provide an accurate estimate of Nm.

Neutrality: Our method assumes the selective equivalence of the genes being sampled. This assumption is needed because we assume that the probability of every pair of genes in a deme coalescing in any generation is the same. It is not entirely clear what effect selection would have on our method. One possible effect would be on reconstructing the phylogeny. Some methods for reconstructing phylogenies may yield erroneous phylogenies if rates of evolution in some lineages are much higher than on others (FELSENSTEIN 1988).

If the reconstructed phylogeny is accurate, the principal effect of selection would probably be on the effective population size. If an advantageous mutation occurs and sweeps through a population, it tends to reduce effective population size. Theoretical analysis shows that the effect of "hitchhiking" can be substantial if there is no recombination (MAYNARD SMITH and HAIGH 1974). On the other hand, if some genes are deleterious but are maintained by recurrent mutation, effective population size would also be smaller because deleterious genes would tend to have coalesced more recently in the past than the neutral

theory would suggest. Hence, two obvious kinds of selection to consider tend to reduce effective population size and reduce estimates of Nm. Until more is known about selection on mitochondrial DNA, it seems impossible to say how important this source of bias might be.

Other population structures: The island model that we have considered represents the extreme in long distance gene flow. Migration is most effective in this population structure and would be less so in other population structures. Therefore, if the actual pattern of migration were more restricted it would require higher levels of gene flow to result in the same degree of mixing. Hence, the estimate of Nm obtained with our method is a minimum estimate, although by how much we do not know.

Low levels of gene flow: If there are low levels of gene flow, our method is not well suited to making accurate estimates of Nm. The reason, which is clear from Figure 2, is that $\bar{s}$ and the entire distribution of s is strongly dependent on Nm when Nm is small ($Nm < 1.0$). In a study of mitochondrial DNA, only a single value of s can be obtained and if Nm is small that value is likely to be small as well. Because s is necessarily an integer value, there are few estimates of Nm that are possible and the confidence limits on each estimate are relatively large. In particular, if $s = 1$, then the estimate of Nm is 0.0. SLATKIN (1989) and TAKAHATA and SLATKIN (1989) consider this case in more detail and SLATKIN (1989) provides a way to place an upper bound on the estimate of Nm that is consistent with such data.

Recombination: We have assumed that there has been no recombination among genes sampled. That assumption appears to be valid for animal mitochondrial DNA, both mitochondrial and chloroplast DNA from plants and from some portions of sex chromosomes. Only in the absence of recombination does the phylogeny of the genes sampled have meaning. Recombination would mix parts of genes with different ancestry so it would be inappropriate to assume there is a single ancestor for a pair of genes. HUDSON (1983) and HUDSON and KAPLAN (1985) discusses this problem in greater detail. Nevertheless, it is possible to construct a phenogram representing the degrees of similarity of genes and apply our method to the phenogram as if it were a phylogeny. At the present time, we do not know whether this would result in an accurate estimate of Nm.

Nuclear *vs.* mitochondrial and chloroplast genomes: Currently, the analysis of nuclear genes is done largely using electrophoretic surveys of numerous loci, while mitochondrial and chloroplast genomes are examined using restriction enzymes or sequencing. Although population surveys of DNA sequence variation are now being carried out for a few nuclear loci in a few species of Drosophila it will probably be some time before large samples of loci from different populations of the same species will be sequenced. The new methods for rapidly sequencing small portions of DNA are, for technical reasons, being applied primarily to portions of mitochondrial and chloroplast genomes, making their use equivalent to a high resolution restriction enzyme analysis.

It is reasonable to suppose that the much greater detail available about mitochondrial and chloroplast genomes would always provide more information about population structure, but the results presented here suggest that is not necessarily so. In fact, we found that confidence intervals obtained by applying our method to the phylogeny of a single gene are not smaller than those obtained using F_{ST} with electrophoretic data. The reason is that a single gene is subject to a variety of accidents in its history because of the intrinsically stochastic nature of population genetic processes. Even knowledge of the complete DNA sequence of each gene cannot overcome that fact. Although electrophoresis offers much less resolution, it does allow the analysis of large numbers of more or less independent nuclear loci, making it possible to average over accidents in their history. We are not suggesting that electrophoresis is preferable for analyzing population structure but at the present time it is not worse either. Ideally both methods would be used because of the possibility that nuclear and extranuclear genes might be subject to somewhat different evolutionary forces. Eventually, rapid sequencing will be used on a variety of nuclear genes making the present choice unnecessary.

CONCLUSION

We have shown that our method for estimating Nm from the phylogenies of genes provides estimates that are nearly as accurate as other indirect methods that have been applied to allozyme data. As is the case with other indirect methods, we cannot distinguish the effects of ongoing gene flow from the effects of historical association of populations. An estimate of Nm obtained using our method must be interpreted as meaning there is ongoing gene flow in a collection of populations at equilibium, or historical association of those populations, or some mixture. If Nm is found to be substantially greater than one, either there is enough gene flow at present to prevent substantial divergence of neutral genetic loci or the recent historical association of the populations sampled.

One feature of our method is that it can be applied to any inherited component for which there is no recombination. If samples of DNA from both mitochondria and the nonrecombining portion of the Y chromosome were available from the same species of

mammals, then separate estimates of male and female dispersal could be obtained.

The use of our method and any method that depends on phylogenies of genes requires a shift in perspective on within-species variation. Allozyme data followed in the tradition of classical population genetics in providing estimates of gene frequencies. Such data could then be examined using classical methods such as F_{ST}. There is some tendency to regard data based on restriction site polymorphisms and even DNA sequences in the same way. It is true that frequencies of haplotypes can be found and then analyzed using methods suitable for analyzing gene frequencies. Although this approach is not incorrect, it does not make full use of information in the data. Furthermore, when direct sequencing becomes easier and it is found that most or all segments of DNA sampled are unique, the gene frequency approach breaks down. In contrast, a phylogenetic approach to the analysis of genetic variation becomes more powerful with greater resolution of the data.

This research was supported in part by a U.S. Public Health Service grant GM40282 to M.S., by the Miller Foundation for Basic Research and by a Canadian NSERC post-doctoral fellowship to W.P.M. We thank S. A. FRANK and N. TAKAHATA for helpful comments on an earlier version of this paper and for J. C. AVISE for providing additional information about his data.

LITERATURE CITED

AVISE, I. C., I. ARNOLD, R. M. BALL, E. BERMINGHAM, T. LAMB, I. E. NEIGEL, C. A. REEB and N. C. SAUNDERS, 1987 Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Annu. Rev. Ecol. Syst. **18:** 489–522.

BERMINGHAM, E., and J. C. AVISE, 1986 Molecular zoogeography of freshwater fishes in the southeastern United States. Genetics **113:** 939–965.

FARRIS, J. S., 1970 Methods for computing Wagner trees. Syst. Zool. **18:** 83–92.

FELSENSTEIN, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:** 521–565.

FITCH, W. M., 1971 Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20:** 406–416.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., and N. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

KINGMAN, J. F. C., 1982a The coalescent. Stochast. Proc. Appl. **13:** 235–248.

KINGMAN, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

MADDISON, W. P., 1989 Reconstructing character evolution on polytomous evolutionary trees. Cladistics (in press).

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitchhiking effect of a favorable gene. Genet. Res. **23:** 23–35.

SLATKIN, M., 1985 Gene flow in natural populations. Annu. Rev. Ecol. Syst. **16:** 393–430.

SLATKIN, M., 1989 Detecting small amounts of gene flow from phylogenies of alleles. Genetics **121:** 609–612.

SLATKIN, M., and N. H. BARTON, 1989 A comparison of three indirect methods for estimating the average level of gene flow. Evolution **43:** 1349–1368.

TAKAHATA, N., 1988 The coalescent in two partially isolated diffusion populations. Genet. Res. **52:** 213–222.

TAKAHATA, N., and M. SLATKIN, 1989 Geneology of neutral genes in two partially isolated populations. Theor. Popul. Biol. (submitted for publication).

TAVARÉ, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetic models. Theor. Popul. Biol. **26:** 119–164.

Communicating editor: B. S. WEIR