Estimation of Levels of Gene Flow From DNA Sequence Data

Richard R. Hudson,* Montgomery Slatkin† and Wayne P. Maddison‡

**Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717, †Department of Integrative Biology, University of California, Berkeley, California 94720, and ‡Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721*

Manuscript received October 18, 1991
Accepted for publication June 16, 1992

ABSTRACT

We compare the utility of two methods for estimating the average levels of gene flow from DNA sequence data. One method is based on estimating F_{ST} from frequencies at polymorphic sites, treating each site as a separate locus. The other method is based on computing the minimum number of migration events consistent with the gene tree inferred from their sequences. We compared the performance of these two methods on data that were generated by a computer simulation program that assumed the infinite sites model of mutation and that assumed an island model of migration. We found that in general when there is no recombination, the cladistic method performed better than F_{ST} while the reverse was true for rates of recombination similar to those found in eukaryotic nuclear genes, although F_{ST} performed better for all recombination rates for very low levels of migration ($Nm = 0.1$).

IN this paper, we will compare the performance of two methods for estimating the average level of gene flow from DNA sequence data. We will be concerned with a data set that consists of a list of sequences for homologous segments of DNA from individuals sampled from different geographic locations. We assume that in each chromosome sampled either the region of interest has been completely sequenced or that presence or absence of numerous polymorphic restriction sites are known. There is already an abundance of such data available for mitochondrial DNA (AVISE *et al.* 1987) and some data for nuclear genes, primarily in Drosophila (*e.g.*, KREITMAN 1983; RILEY, HALLAS and LEWONTIN 1989; SCHAEFFER and MILLER 1991). One question that arises is how these data can be used most effectively to infer something about population structure.

There are currently two methods available for using sequence data to estimate the average level of gene flow. One method treats each polymorphic site as a separate locus and then estimates F_{ST} from the frequencies of alleles at each locus in different geographic locations. Calculating F_{ST} for this kind of data was first suggested by NEI (1982) and later by TAKAHATA and PALUMBI (1985) and by LYNCH and CREASE (1990). We will discuss later the differences between these ways of estimating F_{ST} from sequence data. From the estimate of F_{ST}, the level of gene flow as measured by the product Nm can be computed from WRIGHT's (1951) result for haploid organisms in an island model of population structure. For this model, WRIGHT found that

$$F_{ST} \approx \frac{1}{1 + 2Nm} \tag{1}$$

where N is the number of individuals in each subpopulation and m is the fraction of migrants in each subpopulation in each generation. An estimator of Nm, which we will denote $\langle Nm \rangle_F$, can be obtained by solving (1) for Nm,

$$\langle Nm \rangle_F = \frac{1}{2}\left(\frac{1}{F_{ST}} - 1\right), \tag{2}$$

where F_{ST} in this equation is actually an estimate of F_{ST}. Equations 1 and 2 are appropriate for haploid organisms or mitochondrial DNA which can be treated as a haploid genome. For mitochondrial DNA and assuming that inheritance is strictly maternal, N in Equation 1 is the effective number of females in each subpopulation. For diploids the 2 in Equation 1 is replaced by a 4, and N in this case is the effective number of diploid individuals in each subpopulation. We note that estimates of F_{ST} can be expressed in terms of average divergence between pairs of sequences within subpopulations and average divergence between pairs of sequences randomly drawn from the whole population (SLATKIN 1991). For this reason we refer to the method of estimating Nm based on F_{ST} as a pairwise method.

SLATKIN and MADDISON (1989) introduced a second method to estimate Nm from sequence data. With their method, the first step is to use the sequence data to infer the gene tree of the samples. Then a parsimony criterion is used to obtain s, the minimum number of migration events consistent with the gene

tree and the geographic locations from which the samples were taken. SLATKIN and MADDISON (1989) carried out extensive simulations to show that the value of s could be used to estimate Nm. The estimator based on s will be denoted $\langle Nm \rangle_s$.

In this paper we will apply both of these methods to simulated data to determine how well each performs under known conditions. We will be particularly concerned with the effects of recombination.

METHODS

Simulation program: The data to which the two methods were applied were produced by a computer program that first generates the genealogy of a sample of genes, and then places mutations randomly on the genealogy to produce a sample. The program generates data under a Wright-Fisher neutral model with population subdivision. We have assumed a finite island model, in which each of d subpopulations receives equal numbers of migrants from each other subpopulation. In addition, it is assumed that recombination can occur at any of a large number of equivalent sites in the genetic region being examined. All mutations are assumed to occur at previously unhit sites, *i.e.*, we assumed an infinite-sites model. To generate samples under this model two previously described algorithms, that of STROBECK (1987) and that of HUDSON (1983), are combined. STROBECK (1987) has described a method for generating gene genealogies under the island model. HUDSON (1983) described how to generate genealogies under models in which recombination is possible between any of a large number of sites. To incorporate both recombination and migration is a straightforward extension of these methods. Copies of the program are available on request. It is important to note that we are assuming a steady state model with continuous levels of gene flow, and that this is quite distinct from models in which subpopulations are now completely isolated but were derived from ancestral populations at some time in the past.

In all the simulations reported here, it was assumed that the number of subpopulations, d, is 10, and that samples of size 16 were taken from two of the subpopulations. It was assumed that a homologous genetic region was sequenced in each of the 32 sampled genes and that a total of 128 polymorphic sites were found in the sample. Four different migration rates were used, Nm = 0.1, 1.0, 5.0 and 10.0. Five different recombination rates were used, namely, Nr = 0.0, 2.0, 4.0, 8.0 and 16.0. The levels of recombination considered here are well within the range estimated for natural populations. For each combination of parameter values, 1000 replicate samples were generated and hence 1000 estimates of Nm were obtained using each of two methods.

Data analysis: For each sample, F_{ST} was estimated by

$$\langle F_{ST} \rangle = 1 - \frac{H_w}{H_b} \tag{3}$$

where H_w is mean number of differences between different sequences sampled from the same subpopulation, and H_b is the mean number of differences between sequences sampled from the two different subpopulations sampled. In other words, for each generated sample in our simulations, H_w is the average of the 240 $\left(= 2\binom{16}{2}\right)$ pairwise comparisons of sequences within subpopulations. Hb is the average of 256 ($= 16^2$) pairwise comparisons of sequences from the two different subpopulations sampled. This estimator of F_{ST} is numerically identical to $\hat{\theta}$ of WEIR and COCKERHAM (1984) for the case of random union of gametes with equal sample sizes from each subpopulation, and where the information from each polymorphic site is combined as WEIR and COCKERHAM recommend for combining information from different loci. Our estimator is also almost the same as N_{ST} of LYNCH and CREASE (1990), differing only in not performing a Jukes-Cantor correction. We do not apply a Jukes-Cantor correction because we assume an infinite-sites model which does not require any correction for multiple hits. Our estimator is slightly different from NEI's (1982) γ_{ST} because in computing H_w our estimator does not include a comparison of a sampled sequence with itself while γ_{ST} does. Our estimator is also slightly different from that of TAKAHATA and PALUMBI (1985) which was designed for application to differences in the number of polymorphic restriction sites. TAKAHATA and PALUMBI assume that a restriction site must be detected as polymorphic before it is counted. That assumption leads to a slight difference between their estimator and Equation 3. Both NEI's and TAKAHATA and PALUMBI's estimators differ from (3) by terms of order $1/n$ where n is the number of sequences sampled from each subpopulation.

Using Equations 3 and 2, one finds the following expression for our estimator:

$$\langle Nm \rangle_F = \frac{1}{2} \frac{H_w}{H_b - H_w}. \tag{4}$$

The statistical properties of $\langle Nm \rangle_F$ as estimated from our simulations are shown in Table 1. H_w is an estimate of the average divergence time of pairs of genes sampled from within a subpopulation and H_b is an estimate of the average divergence time of genes sampled from different subpopulations. If H_b is less than or equal to H_w, then $\langle Nm \rangle_F$ is negative or infinity.

TABLE 1

Comparison of statistical properties of $\langle Nm \rangle_F$ and $\langle Nm \rangle_s$

Nr	P_{def}	Mean	(Variance)	$C(0.025)$	$C(0.05)$	$C(0.5)$	$C(0.95)$	$C(0.975)$
$Nm = 0.1$								
0.0	0.999	0.340	(1.23)	0.0026	0.0041	0.094	1.2	1.6
		0.157	(0.27)	0.0	0.0	0.0	0.35	0.75
2.0	1.000	0.17	(0.094)	0.0048	0.0069	0.091	0.54	0.73
		0.077	(0.036)	0.0	0.0	0.0	0.35	0.35
4.0	1.000	0.17	(0.044)	0.0053	0.0073	0.104	0.49	0.64
		0.066	(0.051)	0.0	0.0	0.0	0.35	0.35
8.0	1.000	0.13	(0.017)	0.010	0.016	0.1030	0.36	0.49
		0.036	(0.016)	0.0	0.0	0.0	0.35	0.35
$Nm = 1.0$								
0.0	0.976	4.96	(754)	0.19	0.29	1.41	13.0	22.1
		1.47	(2.9)	0.0	0.35	1.3	3.24	4.95
2.0	0.997	1.89	(6.63)	0.29	0.37	1.28	5.20	7.19
		1.39	(1.58)	0.0	0.35	1.30	3.24	4.95
4.0	1.000	1.60	(2.40)	0.34	0.43	1.24	3.88	4.97
		1.22	(1.34)	0.0	0.0	0.75	3.24	3.24
8.0	1.000	1.41	(1.06)	0.41	0.49	1.16	3.20	3.92
		1.03	(1.10)	0.0	0.0	0.75	3.24	3.24
16.0	1.000	1.31	(0.553)	0.476	0.576	1.13	2.63	3.10
		0.71	(0.544)	0.0	0.0	0.35	1.99	1.99
$Nm = 5.0$								
0.0	0.781	31.2	(32,600)	0.90	1.23	5.80	73.1	162
		11.8	(346)	1.3	1.99	4.95	36.6	73.04
2.0	0.948	15.6	(2,810)	1.59	1.92	6.48	42.9	64.7
		10.0	(207)	1.3	1.99	4.95	36.57	36.57
4.0	0.968	43.8	(748,000)	1.59	2.11	6.21	40.3	94.7
		9.62	(168)	1.3	1.99	4.95	36.57	36.57
8.0	0.989	10.7	(666)	2.03	2.31	6.09	27.6	41.7
		8.33	(158)	1.3	1.3	4.95	36.57	36.57
16.0	0.996	8.91	(171)	2.27	2.66	6.23	22.2	31.9
		7.48	(116)	1.3	1.3	4.95	15.88	36.57
$Nm = 10.0$								
0.0	0.628	41.7	(39,300)	1.56	2.04	8.94	114	238
		29.6	(1,640)	3.24	3.24	15.88	73.04	146.08
2.0	0.836	60.4	(342,000)	2.54	3.21	11.9	151	226
		24.98	(1,590)	1.99	3.24	15.88	73.04	73.04
4.0	0.860	36.4	(27,400)	3.22	3.66	11.4	117	198
		25.46	(1,440)	3.24	3.24	15.88	73.04	146.08
8.0	0.909	45.7	(63,700)	3.48	4.21	12.0	124	221
		22.51	(885)	1.99	3.24	15.88	73.04	146.08
16.0	0.941	31.1	(28,600)	4.33	4.78	11.9	66.0	109
		20.10	(857)	1.99	3.24	7.93	73.04	73.04

For each value of Nm and Nr, there are two lines, the top line is for $\langle Nm \rangle_F$ and the second line is for $\langle Nm \rangle_s$. P_{def} is an estimate of the probability of the estimator $\langle Nm \rangle_F$ being defined. (If F_{st} is less than or equal to 0.0, the estimator is considered undefined.) Everything to the right of P_{def} is conditional on the estimator being defined. $C(x)$ is the estimated value for which the probability of the estimator being less than $C(x)$ is x. ($C(0.5)$ is an estimate of the median.) All estimates are based on 1,000 samples.

When this occurred in a replicate, we say that the estimator is undefined for that sample.

To use SLATKIN and MADDISON's (1989) cladistic method, we began by inferring a gene tree of the samples using the computer program PAUP (SWOFFORD 1990) which applies a parsimony algorithm to finding the best gene tree. A fast heuristic algorithm was used: simple addition sequence with no branch swapping. (It seems unlikely that a more thorough algorithm would have made a difference; robustness is suggested by pilot studies with UPGMA trees that gave qualitatively similar results.) If there is recombination, then the history of the sample of genes cannot in general be represented by a single gene tree. PAUP can, however, still be used to infer a tree. Given the inferred gene tree, we then used the algorithm described by SLATKIN and MADDISON for finding s, the minimum number of migration events. From that value of s and the simulation results of SLATKIN and MADDISON (1989, Table 1) we estimated Nm. Because s can take only integer values, the estimates of Nm that are possible for given sample sizes also take only discrete values. For 16 genes sampled from each of two populations, we interpolated the values from Table 1 of SLATKIN and MADDISON (1989) to obtain the estimates of Nm shown in Table 1 here. These

were the values used in analyzing the simulation results.

RESULTS

Table 1 shows estimated mean, variances and some percentiles of the two estimators, $\langle Nm \rangle_s$ and $\langle Nm \rangle_F$, for several migration rates and recombination rates. The properties of the estimator $\langle Nm \rangle_F$ are summarized first. The median of this estimator is within 40% of the true value for all migration rates and recombination rates examined, and for most parameter values the median is within 20%. When recombination rates are high the mean is also close to the true value, except at the highest level of migration. However, when recombination rate is low and/or the migration rate is high, the mean can be considerably higher than the true value. In these cases, the variance is also very large and the estimator is frequently undefined. Figure 1 shows that, with low recombination, the distribution of the estimator is highly skewed with a large tail to the right, indicating the substantial probability of very large values of the estimator. These large values of the estimator are clearly due to cases where the denominator of the right hand side of (4) is near zero. Figure 1 also shows that recombination has the effect of shrinking the large tail to the right in the distribution of the estimator, reducing the variance and the bias of the estimator.

The estimator based on s has a discrete distribution. When migration rates are low the estimator is zero a large fraction of the time. Recombination has a less drastic effect on this estimator, reducing the mean and the variance to some extent. Except when migration rate is very low, recombination also improves this estimator, reducing the bias and variance.

Comparing the two estimators, we see that if migration rate is very low, the estimator based on F_{ST} is clearly superior. With moderate to high migration and low recombination, the estimator based on s is clearly better, having less bias and much lower variance. With higher levels of recombination, the situation is less clear. For moderate migration and high recombination, the two estimators are statistically very similar, with the estimator based on F_{ST} being perhaps slightly better. With higher levels of migration, and with high recombination, Figure 1 shows that the two estimators have very similar distributions. However, with low probability the estimator based on F_{ST} can take quite large values, which leads to a larger bias and variance than the estimator based on s.

DISCUSSION

We can understand these results by considering the kinds of information used by each of these methods and the effects of recombination. First, to understand why $\langle Nm \rangle_F$ works better than $\langle Nm \rangle_s$ when Nm is small consider the following argument. If the migration rate is sufficiently small, then most of the time all the sequences coalesce in each subpopulation before any migration events occur (as one proceeds back in time tracing the history of the sampled sequences). That is, samples from the different localities form clades connected by relatively long branches (SLATKIN 1989). In this case, only one migration event will be required in a parsimonious reconstruction of the history of the sampled sequences. This is a migration event bringing an ancestor of the sample from one locality to the locality of the other sample. One is the minimum number of migration events possible for samples from two localities. Therefore, for the estimate based on s, there is no way to distinguish small rates of migration from very small rates of migration, they both usually require only one migration event. On the other hand, F_{ST} depends on the length of the branch that connects the two clades, which is dependent on the value of Nm, especially for Nm small (SLATKIN 1991). The estimate, $\langle Nm \rangle_F$, utilizes that dependence and therefore outperforms $\langle Nm \rangle_s$ for Nm small.

We can also see, at least in part, why the estimator, $\langle Nm \rangle_F$, is better with higher levels of recombination. If there is no recombination, then each sample of genes has a simple unique gene tree which differs from replicate to replicate. In calculating F_{ST}, only some information from that gene tree is extracted, namely the average divergence times of genes sampled from the same and from different populations. As discussed by FELSENSTEIN (1992) and others, the problem with using average divergence times in such cases is that they are very strongly effected by the deepest branch in the gene tree. The cladistic method of SLATKIN and MADDISON (1989), on the other hand, uses more information from the gene tree, namely the relationship between its topology and the geographic locations from which the genes are sampled. That method still does not use all the information because it ignores branch lengths.

With recombination, there is no longer a single simple gene tree of the genes sampled. Instead, each non-recombined segment of DNA has its own, partially independent, gene tree. The F_{ST} method benefits from the fact that it averages over events at different sites which have different and partially independent trees. It is still based on estimates of average divergence times of genes but the estimates of divergence times will be less variable because they are in effect averaged over different realizations of the process that generates trees. The cladistic method on the other hand depends on inferring a single gene tree and with some recombination the inferred gene tree cannot represent the actual history of the sampled genes. Instead it represents some kind of average of the gene trees of the nonrecombined segments. For

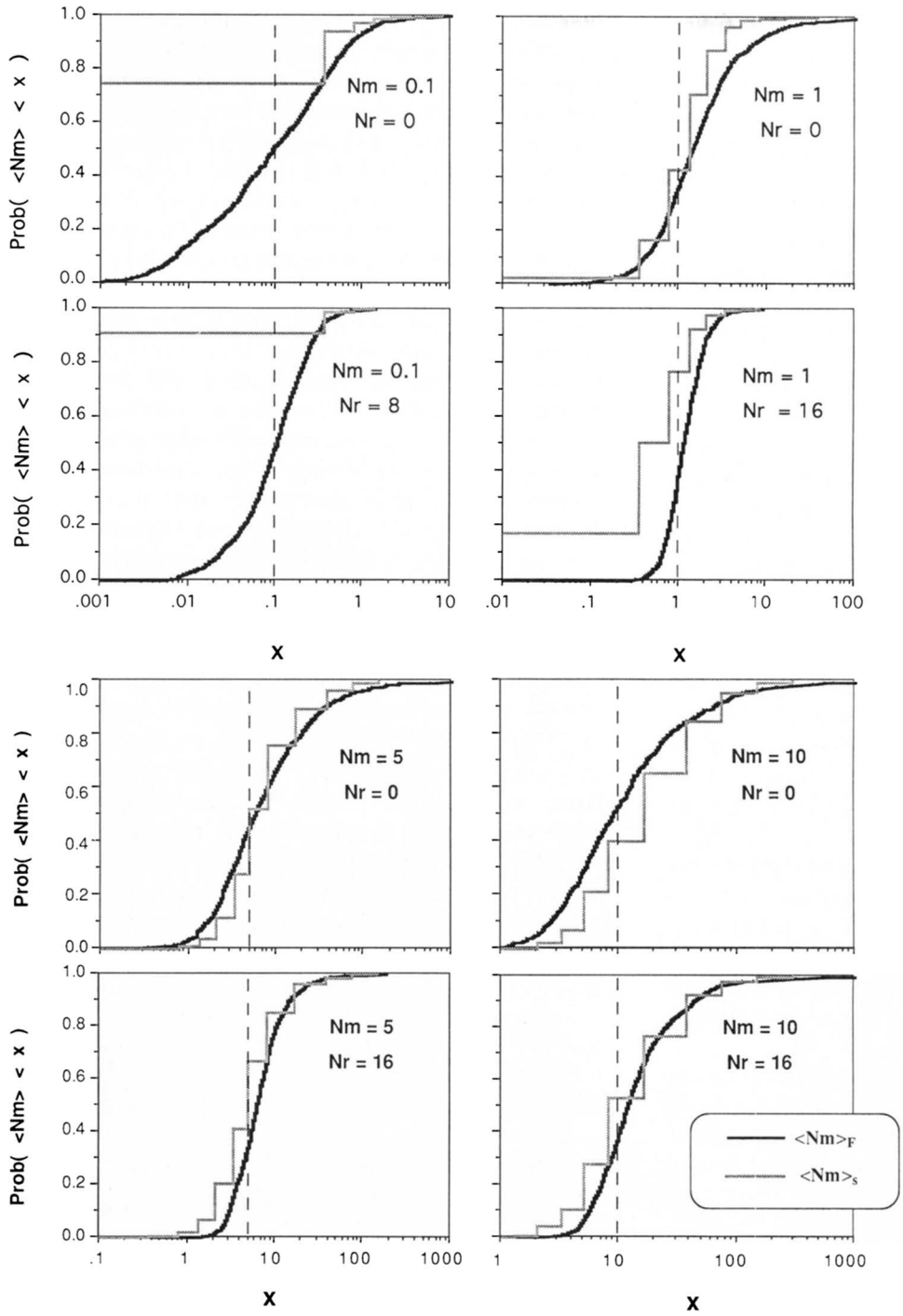

FIGURE 1.—Estimated cumulative probability distributions of the two estimators, $\langle Nm \rangle_F$ and $\langle Nm \rangle_s$. The dashed vertical line in each plot indicates the true value of Nm. These distributions are based on the same simulations used to generate Table 1.

the cladistic method, there is no benefit from this averaging and apparently little increase in accuracy because the estimate of Nm is still based on a single number.

Recombination could be accounted for in the cladistic method by identifying the locations of the recombination events using HEIN's (1990) or some other method. Then an estimate of Nm could be obtained from each nonrecombined segment and those estimates averaged. The cladistic method will almost certainly perform better if it is used this way because we have already shown that it will perform better for each nonrecombined segment. However, it seems likely that it will take a rather large amount of data for this modification of the cladistic method to be practicable.

The results in Table 1, show that the estimator, $\langle Nm \rangle_F$, is biased even in the best of circumstances. This is, in large part, due to the large variance of the denominator in the expression on the right hand side of (4). But even when the variance of the denominator is small some bias remains. This can be seen by examining the ratio of the expectations of the numerator and denominator of the right hand side of (4). Using

the expectation of H_w and H_b under the neutral island model (LI 1976a), we find that ratio of the expectations is $[d/(d-1)]Nm$, where d is the number of subpopulations. (Note that d is the actual number of subpopulations, not the number of subpopulations sampled.) Thus, for our simulations where d equals 10, a bias of about 10% is expected even under the best conditions. If d is known, better estimates might be obtained using the result of LI (1976b),

$$F_{ST} \approx \frac{1}{1 + \left(\frac{d}{d-1}\right)^2 2Nm} \tag{5}$$

instead of (1) and estimating F_{ST} by

$$\langle F_{ST} \rangle_d = 1 - \frac{H_w}{\frac{1}{d}H_w + \frac{d-1}{d}H_b}. \tag{6}$$

If one solves (5) for Nm and replaces F_{ST} by the estimate in (6), one arrives at the estimator

$$\begin{aligned} \langle Nm \rangle &= \left(\frac{d-1}{d}\right)^2 \frac{1}{2}\left(\frac{1}{\langle F_{ST} \rangle_d} - 1\right) \\ &= \frac{d-1}{d}\frac{1}{2}\frac{H_w}{H_b - H_w}. \end{aligned} \tag{7}$$

which differs from $\langle Nm \rangle_F$ by the factor $(d-1)/d$. Under the finite island model, the ratio of the expectations of the numerator and denominator of the right hand side of (7) is Nm. For the situation that we considered, where d equals 10, the bias of $\langle Nm \rangle$ is somewhat less than that of $\langle Nm \rangle_F$ and the variance is about 20% less. Thus when the number of subpopulations is known and small, the estimator $\langle Nm \rangle$ would appear to be better than $\langle Nm \rangle_F$.

Finally we note that, everything else being equal, genetic regions with recombination provide better estimates of Nm than regions without recombination. This is true for both estimators, although for $\langle Nm \rangle_s$ the effect of recombination is not so great.

CONCLUSIONS

We conclude that available methods can be applied to within-species DNA sequence data to provide reasonably accurate estimates of the average level of gene flow, as measured by Nm. However, if Nm is large then the distributions of the estimators are very skewed toward large values. In this case, the estimators can be very biased with high variance, though the median values of the estimators remain close to the true values. If the region sequenced has little or no recombination, as in the case of mitochondrial DNA in animals, then SLATKIN and MADDISON's (1989) is likely to be more accurate than a method based on estimating F_{ST} from polymorphic sites. If there are very low levels of gene flow, the cladistic method would probably yield a zero estimate of Nm, which would mean that samples from each geographic location formed a single clade in the unrooted gene tree. In that case, F_{ST} would give a nonzero estimate of Nm that would give a better although somewhat biased estimate of how low the level of gene flow is.

For higher levels of recombination, as found in most parts of the nuclear genome, the F_{ST} estimator is generally better. The difference between F_{ST} and the cladistic method is surprisingly small, given how badly the assumptions underlying the cladistic method are violated when recombination rates are high. The cladistic method does not tend to be biased upwards as much as the F_{ST} method.

We do not think that either of the methods described here are the best possible, only that they are available now and are relatively easy to use. The method based on estimating F_{ST} uses only the average pairwise differences between sequences and does not make any use of the topological structure of the gene tree or trees that more completely describe the data. The cladistic method makes use only of the topological structure of a single inferred gene tree and not the branch lengths or any variability among gene trees describing a single data set.

This study was supported in part by grants from National Institutes of Health to R.R.H. (GM42447) and M.S. (GM40282) and by a National Sciences and Engineering Research Council of Canada postdoctoral fellowship to W.P.M.

LITERATURE CITED

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

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise estimation as compared to phylogenetic estimation. Genet. Res. **59:** 139–147.

HEIN, J., 1990 Reconstructing evolution of sequences subject to recombination using parsimony. Math. Biosci. **98:** 185–200.

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

KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

LI, W.-H., 1976a Distribution of nucleotide differences between two randomly chosen cistrons in a subdivided population: the finite island model. Theor. Popul. Biol. **10:** 303–308.

LI, W.-H., 1976b Effect of migration on genetic distance. Am. Nat. **110:** 841–847.

LYNCH, M., and T. J. CREASE, 1990 The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. **7:** 377–394.

NEI, M., 1982 Evolution of human races at the gene level, pp. 167–181 in *Human Genetics, Part A: The Unfolding Genome*, edited by B. BOHHE-TAMIR, P. COHEN and R. N. GOODMAN. Alan R. Liss, New York.

RILEY, M. A., M. E. HALLAS and R. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the *Xdh* locus in Drosophila pseudoobscura. Genetics **123:** 359–369.

SCHAEFFER, S. W., and E. L. MILLER, 1991 Nucleotide sequence analysis of *Adh* genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. Proc. Natl. Acad. Sci. USA **88:** 6097–6101.

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

SLATKIN, M., 1991 Inbreeding coefficients and coalescence time. Genet. Res. **58:** 167–175.

SLATKIN, M., and W. P. MADDISON, 1989 A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics **123:** 603–613.

STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision. Genetics **117:** 149–153.

SWOFFORD, D. L., 1990 PAUP (Phylogenetic Analysis Using Parsimony) version 3.0h. Illinois Natural History Survey, Champaign.

TAKAHATA, N., and S. R. PALUMBI, 1985 Extranuclear differentiation and gene flow in the finite island model. Genetics **109:** 441–457.

WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating *F*-statistics for the analysis of population structure. Evolution **38:** 1358–1370.

WRIGHT, S., 1951 The genetical structure of populations. Ann. Eugenics **15:** 323–354.

Communicating editor: B. S. WEIR