Optimal Sequencing Strategies for Surveying Molecular Genetic Diversity

Anna Pluzhnikov* and Peter Donnelly*,†

*Department of Statistics and †Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received February 29, 1996
Accepted for publication July 22, 1996

ABSTRACT

Two commonly used measures of genetic diversity for intraspecies DNA sequence data are based, respectively, on the number of segregating sites, and on the average number of pairwise nucleotide differences. Expressions are derived for their variance in the presence of intragenic recombination for a panmictic population of fixed size that is at neutral equilibrium at the region sequenced. We show that, in contrast to the slow decrease in variance with increasing sample size, if the recombination rate is nonzero, the asymptotic rate of decrease of variance with increasing sequence length, for fixed sample size, is quite rapid. In particular, it is close to that which would be obtained by sequencing independent chromosome regions. The correlation between measures of diversity from linked regions is also examined. For a given total number of bases sequenced in a particular region, optimal sequencing strategies are derived. These typically involve sequencing relatively few (three to 10) long copies of the region. Under optimal strategies, the variances of the two measures are very similar for most parameter values considered. Results concerning optimal sequencing strategies will be sensitive to gross departures from the underlying assumptions, such as population bottlenecks, selective sweeps, and substantial population substructure.

RECENT advances in molecular technology have made possible large scale surveys of within-population molecular variation at the DNA sequence level. Such surveys may have several aims. One is descriptive: an attempt to characterize the extent of sequence diversity in a particular population. On a more quantitative level, such surveys might be used to make inferences about the underlying evolutionary mechanisms, perhaps through the estimation of relevant parameters, or the testing of competing hypotheses about the nature of the evolutionary and demographic processes.

This paper is motivated by the question of how best to allocate resources in such a project. In examining a particular chromosomal region, there are choices to be made between the number of copies of the region that are sequenced and the number of nucleotides sequenced in each copy. We consider this trade-off in the context of measuring the genetic diversity in the population at the region in question. Specifically, we study the variability of two commonly used measures of diversity, based respectively on the number of segregating sites in the sample and on the average pairwise number of differences between sequences in the sample.

Our analysis applies in the context of a large, equilibrium, panmictic population that has been of constant size throughout its evolution, for which the evolution of the region in question is neutral. The main theoretical results are derived (in the APPENDIX) using coalescent methods. It is important to note that our conclusions do not necessarily apply under other demographic scenarios. Some discussion of the likely effects of, for example, population bottlenecks or geographical population subdivision is given in the final section of the paper. For this reason, the strategies that we find to be optimal in the context of characterizing diversity may not be optimal for the problem of testing between different evolutionary models.

The next section describes the underlying assumptions and the two most commonly used measures of diversity. Their sampling variances for the infinite sites model are well known. We extend these results to include the possibility of intragenic recombination within the region being sequenced and examine its effect on the precision of the estimators for a range of parameter values. We obtain analytic expressions for the sampling variances of the two estimators that are then used to consider the problem of optimal choice of sample size and sequence length within a particular region. [Earlier, FU (1994) has used simulation methods to estimate the sampling variances in the context of a study of different estimators of the mutation rate]. To facilitate decisions about how much sequencing effort to put into a particular region, and how far apart sequenced regions should be so that conclusions from them may be independent, we also give details of the correlation of the estimators from different loci as a function of the distance between the loci. Technical derivations are given in the APPENDIX, which also includes a derivation of the covariance between sample heterozygosities at two linked loci for which there is no intragenic recombination.

Corresponding author: Anna Pluzhnikov, Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. E-mail: besproz@galton.uchicago.edu

Properties of the estimator based on the number of segregating sites follow from the work of KAPLAN and HUDSON (1985), and HUDSON (1983).

BACKGROUND

We consider measures of genetic diversity from a sample of intraspecies DNA sequences. We assume that the data are obtained from a population that has been panmictic and of constant size throughout its evolution, that the evolution in the region from which the sequences are taken has been neutral, and that the population is at equilibrium in that region under the forces of genetic drift and mutation. We denote by N the effective diploid population size. Write n for the number of sequences in the sample, and assume that each sequence consists of the same number of nucleotide sites, which we denote by L.

Let u denote the mutation probability per site per generation, assumed to be constant across the sites under consideration. Define $\theta = 4Nu$, the scaled mutation rate per site. The mutation rate for the whole of the sequenced region is then $\Theta = L\theta$. In the applications we envisage, L will be large, and θ small, so we will adopt the "infinite-sites" assumption that each mutation since the common ancestor in the genealogical history of the sample will have occurred at a different site.

Two common estimators of diversity are based, respectively, on the number of segregating sites in the data and on the average number of nucleotide differences between each pair of sequences.

If S_n denotes the number of segregating sites in the sample of n sequences, the first of these measures (WATTERSON 1975) is

$$\tilde{\Theta}_W = \frac{S_n}{\sum_{i=1}^{n-1} i^{-1}}, \tag{1}$$

for which

$$E(\tilde{\Theta}_W) = \Theta,$$

and

$$\mathrm{Var}(\tilde{\Theta}_W) = \Theta \frac{1}{\sum_{i=1}^{n-1} i^{-1}} + \Theta^2 \frac{\sum_{i=1}^{n-1} i^{-2}}{(\sum_{i=1}^{n-1} i^{-1})^2}.$$

Suppose the sequences are labeled from $\{1, 2, \ldots, n\}$ and write d_{ij} for the number of nucleotide sites at which sequences i and j differ. A diversity measure based on pairwise differences (TAJIMA 1983) is then

$$\tilde{\Theta}_T = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}, \tag{2}$$

the average number of pairwise differences. For this estimator

$$E(\tilde{\Theta}_T) = \Theta,$$

and

$$\mathrm{Var}(\tilde{\Theta}_T) = \Theta \frac{n+1}{3(n-1)} + \Theta^2 \frac{2(n^2+n+3)}{9n(n-1)}.$$

Our interest here is in the behavior of these measures as both n, the number of sequences, and L, the length of the sequences, vary. Changes in L change Θ, the mean of each estimator. To facilitate comparison, we will scale the measures by the inverse of the length of the region sequenced. Define

$$\tilde{\theta}_W = \frac{\tilde{\Theta}_W}{L}, \quad \text{and} \quad \tilde{\theta}_T = \frac{\tilde{\Theta}_T}{L}. \tag{3}$$

These statistics can be thought of, formally, as estimators of θ, the neutral parameter measuring the scaled mutation rate per site. Then

$$E(\tilde{\theta}_W) = E(\tilde{\theta}_T) = \theta,$$

and

$$\mathrm{Var}(\tilde{\theta}_W) = \theta \frac{1}{L\sum_{i=1}^{n-1} i^{-1}} + \theta^2 \frac{\sum_{i=1}^{n-1} i^{-2}}{(\sum_{i=1}^{n-1} i^{-1})^2}, \tag{4}$$

$$\mathrm{Var}(\tilde{\theta}_T) = \theta \frac{n+1}{3L(n-1)} + \theta^2 \frac{2(n^2+n+3)}{9n(n-1)}. \tag{5}$$

The sampling variances (4) and (5) provide a natural quantification of the precision of the measures. As is well known, for fixed sequence length L, this precision does not improve greatly as the sample size n increases. The variance of $\tilde{\theta}_W$ does decrease to 0 as $n \to \infty$, but slowly, at a rate of $1/\log n$. In contrast, the variance of $\tilde{\theta}_T$ converges to the non-zero limit $\theta/(3L) + 2\theta^2/9$, as $n \to \infty$, so that, formally, the estimator $\tilde{\theta}_T$ is not even consistent.

The reason for this behavior is that distinct sequences sampled from the population are not independent. The type of an additional sequence is positively correlated with the sequences already observed, precisely because it is likely to share a considerable portion of its ancestral history with the other sequences. Relatively little additional evolutionary information is thus gained by examining the additional sequence. The nonconsistency of the pairwise difference estimator is a consequence of the fact that it does not even make good use of this limited additional information. For a fuller discussion, see for example DONNELLY and TAVARÉ (1995).

In view of this relatively minor increase in precision with increasing sample size, it is natural to ask whether one would be better off instead by increasing the length of the region sequenced. We will do so, but first we extend the model to allow for intragenic recombination.

THE EFFECT OF RECOMBINATION

Recombination will be modeled by assuming that there is a constant (small) probability, r, of a recombination between any particular pair of adjacent sites in each generation. We write $\rho = 4Nr$ for the scaled recombination rate between adjacent sites.

The sampling variances of the two estimators in the presence of intragenic recombination are derived in the APPENDIX using coalescent methods. In particular,

$$\mathrm{Var}(\tilde{\theta}_W) \approx \frac{\theta}{L\sum_{i=1}^{n-1} i^{-1}} + \frac{\theta^2}{2L^2(\sum_{i=1}^{n-1} i^{-1})^2} \times \sum_{m=1}^{L-1} (L-m)F_n(m\rho) \approx \frac{\theta}{L\sum_{i=1}^{n-1} i^{-1}} + \frac{\theta^2}{2(\sum_{i=1}^{n-1} i^{-1})^2} \times \int_0^1 (1-x)F_n(L\rho x)\,dx, \quad (6)$$

for large L and small θ, where the function $F_n(z)$ is defined to be the covariance function $F(0, 0, n; z)$, which arises as the solution to the recursive system of equations (A7) in the APPENDIX. Except for $n = 2$, this function must be evaluated numerically. The result (6) is effectively due to KAPLAN and HUDSON (1985), see also HUDSON (1983). For the pairwise difference estimator,

$$\mathrm{Var}(\tilde{\theta}_T) \approx \frac{\theta(n+1)}{3L(n-1)} + \frac{4\theta^2}{n(n-1)} \times \left[\frac{L\rho - C + 13}{2(L\rho)^2}\log\left(\frac{(L\rho)^2 + 13L\rho + 18}{18}\right) + \frac{L\rho(2C-13) + 13C - 133}{19.70(L\rho)^2} \times \log\left(\frac{45.70L\rho + 72}{6.30L\rho + 72}\right) - \frac{1}{L\rho}\right], \quad (7)$$

where $C = 2(n^2 + n + 3)$, and here and throughout, log refers to natural logarithms.

The expressions (6) and (7) reduce to (4) and (5), respectively, when $\rho \rightarrow 0$. In the latter case, this is immediate from the integral expression (A14) in the APPENDIX. In the former, it follows from the fact that $F_n(0) = 4(\sum_{i=1}^{n-1} i^{-2})$, since $F_n(0)$ is simply the variance of the total length of the branches in a n-coalescent tree. For particular nonzero ρ and θ, the sampling variances (6) and (7) of the estimators may be evaluated numerically. We illustrate their values below.

It is interesting to compare the results of the analytic expressions (6) and (7) with those based on a simulation study. Table 2 of FU (1994) presents such estimates of the sampling variances of $\tilde{\theta}_W$ and $\tilde{\theta}_T$ for several values of the total recombination rate $R = L\rho$ and total mutation rate $\Theta = L\theta$ per region. For the values of the parameters considered by FU, a comparison of these estimates with our analytic results shows an extremely close agreement.

To gain some insight into the "trade-off" between the increase in precision of the estimators gained through sequencing additional individuals or sequencing additional bases from the same individuals, focus attention on the segregating sites estimator $\tilde{\theta}_W$. Suppose n sequences of length L are available and contrast the following strategies for obtaining additional information:

(i) Sequence n additional copies of the region.
(ii) Sequence an adjoining region of length L from each of the original sequences.

For the moment, ignore recombination, either within or between the sampled regions.

Under the infinite sites assumption, the number of segregating sites is just the total number of mutations on the genealogical tree linking the sampled sequences with their most recent common ancestor. Write T for the total length of this tree for the original n sampled sequences. Coalescent theory (*e.g.*, DONNELLY and TAVARÉ 1995) gives

$$E(T) = 2\sum_{i=1}^{n-1} i^{-1} \approx 2\log n,$$

for n large. Conditional on T, the number of mutations, and hence the number of segregating sites, is Poisson with parameter $\Theta T/2$.

If strategy (i) above is adopted, the effect of sequencing n additional individuals is to increase the total length of the genealogical tree. If the tree is increased by a length T', the additional information derives from the Poisson process (of rate $\Theta/2$) of mutations, run independently of that on the original tree, for a "time" of length T'. Note that $E(T') = 2\sum_{i=n}^{2n-1} i^{-1} \approx 2\log 2$.

In contrast, under strategy (ii), one gains an independent Poisson process of mutations (again of rate $\Theta/2$), this time superimposed on the original tree. In this case, the independent process runs over a time of length T. The gain in precision under each strategy results exactly from this potential to view an independent realization of the mutation process. The extent of the gain is related to the amount of time over which the independent process runs. On average, this additional time will always be greater under strategy (ii) than under strategy (i). (Note that this heuristic argument does not establish that it is always better to adopt strategy (ii). An exact answer to the question can be obtained by evaluating (4) under each scenario.)

In the absence of recombination in the region of interest, the trees describing genealogical history at each site will be identical. In this sense, trees associated with different sites are maximally positively correlated. The effect of recombination is to allow for different parts of the sequenced region to have distinct genealogical trees. As a consequence, the genealogical trees associated with different sites, and indeed the whole evolutionary process, become more independent. It follows that whatever sequencing strategy is adopted, an increase in the recombination rate ρ will increase the

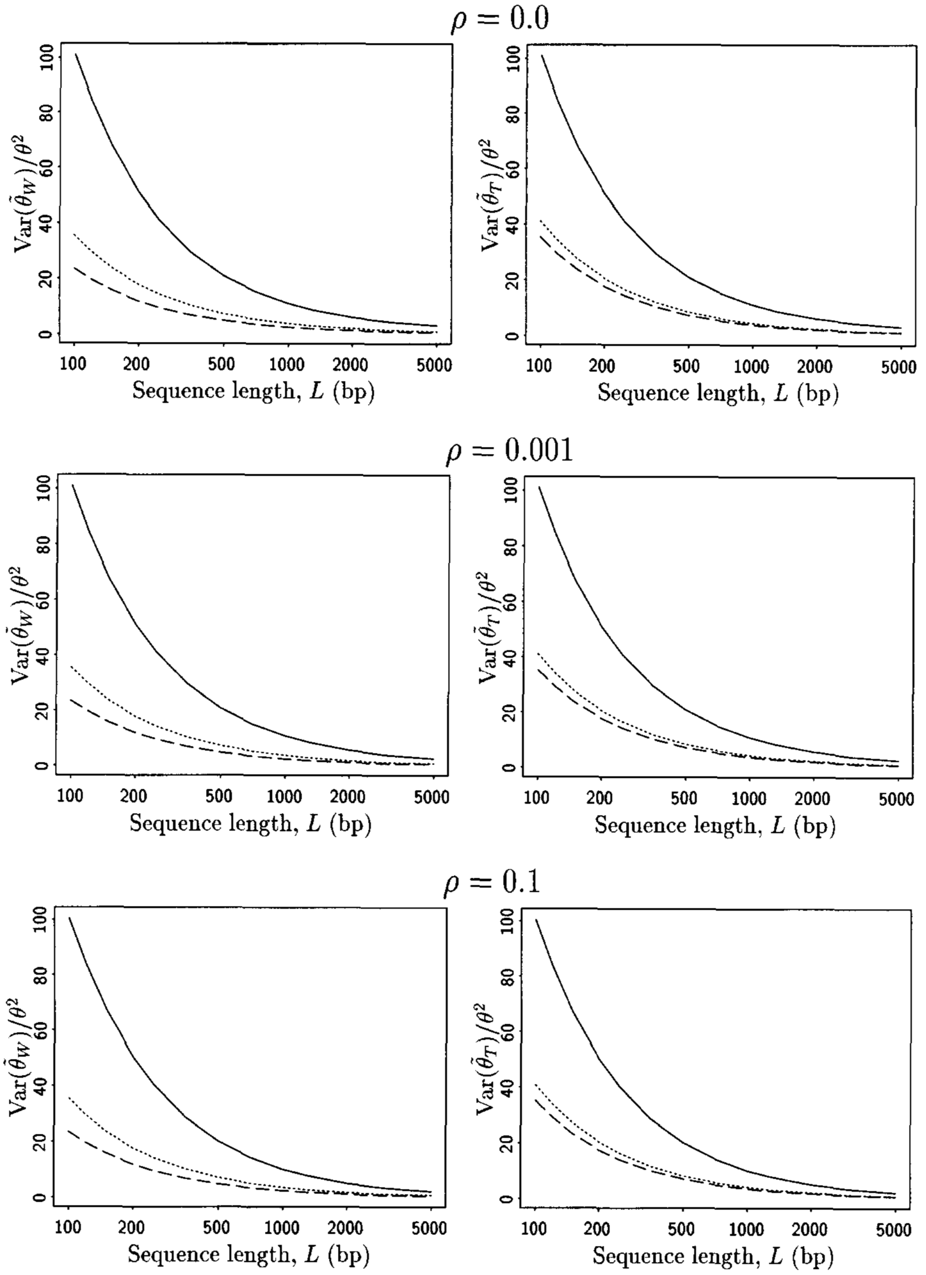

FIGURE 1.—Normalized sampling variance of the estimators $\tilde\theta_W$ (left column) and $\tilde\theta_T$ (right column) as a function of sequence length L, for a fixed value of the scaled mutation rate per site θ = 0.0001, and values of the scaled recombination rate per site ρ = 0.0, 0.001, and 0.1. The lines correspond to the sample size $n = 2$ (——), $n = 10$ (· · ·), and $n = 40$ (– – –).

precision of the estimator $\tilde\theta_W$. Whatever the advantage enjoyed by strategy (ii) above over strategy (i), it will be increased in the presence of recombination.

Figures 1 and 2 show plots of the relative precision of each estimator, for various values of θ, ρ, and n, as a function of L. We also produced plots for these parameter values for ρ = 0.0001 (data not shown). These are very similar to the plots in Figures 1 and 2 corresponding to $\rho = 0$. This relative precision is measured by the squared coefficient of variation, defined as the variance of the estimator divided by the square of θ, the value it is estimating.

The behavior corresponds well with intuition. With other parameters held fixed, the relative precision increases (squared coefficient of variation decreases) as each of n, ρ, θ, or L, is separately increased. (Note the difference in vertical scale for plots corresponding to different values of θ.) The effect on variances of recombination is more marked for larger values of θ. For small θ, it appears that the first term in the variance expressions (6) and (7), which does not depend on ρ, dominates the expression, at least for the values of L in the plots.

For any fixed values of the parameters, the estimator $\tilde\theta_W$ is to be preferred, in the sense of having smaller sampling variance, to the estimator $\tilde\theta_T$. (The estimators coincide for samples of size $n = 2$.) For some parameter values, notably when θ is small, the difference in relative precision between the estimators is, however, small.

We have seen that for fixed θ, ρ, and L, the increase in precision of each estimator as n is increased is very slow. The consistency of $\tilde\theta_W$ in contrast to the nonconsis-

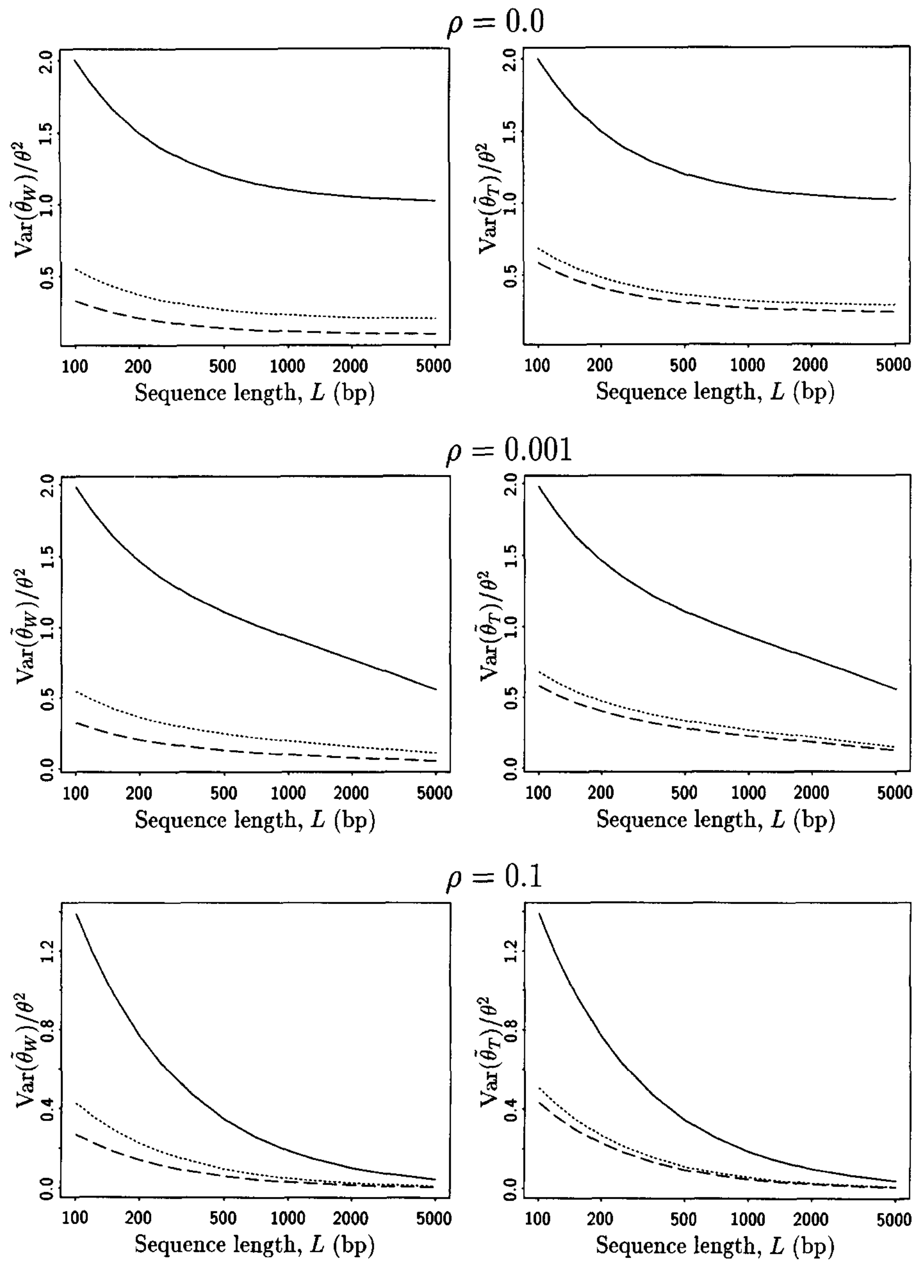

FIGURE 2.—Normalized sampling variance of the estimators $\tilde{\theta}_W$ (left column) and $\tilde{\theta}_T$ (right column) as a function of sequence length L, for a fixed value of the scaled mutation rate per site $\theta = 0.01$, and values of the scaled recombination rate per site $\rho = 0.0$, 0.001, and 0.1. The lines correspond to the sample size $n = 2$ (——), $n = 10$ (· · ·), and $n = 40$ (– – –).

tency of $\tilde{\theta}_T$ is reflected in the fact that for fixed values of the other parameters, the increase in precision caused by a given increase in n is larger for the former than for the latter estimator.

In problems such as this, there are actually two different senses in which one might examine the asymptotic behavior of statistical procedures. The first is the traditional one of fixing the sequence length and increasing the number of sequences sampled. An alternative is to ask about behavior when the number of sequences is held fixed, but the number of bases sequenced is increased.

The plots in Figures 1 and 2 show that for fixed θ, ρ, and n, the sampling variance of the estimators decreases with L. We show in the APPENDIX that provided $\rho > 0$, the expressions (6) and (7) for the variance of the estimators converge to 0 as $L \rightarrow \infty$. For (7) this convergence is at a rate of $(\log L)/L$. The convergence of (6) is at least this fast. (In contrast, if $\rho = 0$, it follows easily from Equations 4 and 5 that the variance of both estimators converges to a nonzero limit as $L \rightarrow \infty$. This effect is particularly clear in the first row of Figure 2.) Thus (for $\rho > 0$) the asymptotic behavior of the estimators as a function of L is more encouraging than their asymptotic behavior as a function of n. The intuition behind this is that recombination acts to generate independence of much of the evolution of different segments of the region, so that as in classical statistical problems, increasing the sequence length generates independent replications of the underlying evolutionary process. In contrast, increasing the sample size gains very little independent replication.

OPTIMAL SEQUENCING STRATEGIES

In practice, an experimenter will typically have flexibility over the choice of both sample size and sequence length. It is thus natural to ask how these quantities should be chosen so as to make best use of limited resources.

Any attempt to formalize this problem requires a specification of the costs incurred by particular experimental strategies, and of the goal of the procedure. We consider the problem in the context of minimizing the variances of the measures $\tilde{\theta}_W$ and $\tilde{\theta}_T$ of genetic diversity.

The cost of sequencing L bases from each of n homologous chromosomes will be a function of L and n. This cost function will in general vary between organisms (depending, for example, on the availability of subjects) and between laboratories (for example, in light of different experimental procedures). We will consider a particular, simple, "cost" function, defined to be the product of L and n. This assumes that the cost of sequencing an additional base is the same, regardless of whether, for example, this involves sequencing an additional individual, or extending the region already sequenced. While this simple cost function is not exact in practice, it may not be completely unrealistic. In addition, the analysis of the problem in this framework may still yield valuable practical insights. It is a straightforward matter to extend the analysis to other, particular, cost functions.

We thus consider the problem of maximizing the precision of the estimators $\tilde{\theta}_W$ and $\tilde{\theta}_T$, when the total number of bases to be sequenced, nL, is fixed. Figure 3 gives plots of the squared coefficient of variation for each estimator, as a function of n, for various values of θ and ρ, when the total number of sites sequenced, nL, is fixed at 10,000. Similar plots for other values of nL display the same broad shape (data not shown). Our plots concern the squared coefficient of variation of the estimators. Since their mean is fixed at θ, exactly the same conclusion would apply for any other monotone function of the sampling variance of the estimators.

For small and moderate values of θ, the minimum of the curves occurs for small values of n. For example, if $\theta = 0.001$, the optimal sequencing strategy (for the range of ρ considered) is to sequence a large region from between three and seven chromosomes, regardless of which estimator is subsequently used. In each case, the measure $\tilde{\theta}_W$ based on the number of segregating sites outperforms $\tilde{\theta}_T$, which is based on the average pairwise difference.

As ρ increases, the optimal sample size decreases and the length of the region sequenced increases. The intuition here is that an increase in the recombination rate increases the degree of independence between the evolution at different sites, so that the gain in precision in extending the region is increased. As θ is increased, the optimal size of the region decreases and the number of chromosomes to be sequenced increases.

For larger θ, and small ρ, the optimal sample size is much larger. For example, for $\theta = 0.01$ and $\rho = 0.001$, the optimal strategy involves sequencing ~25 copies of the region. Note, however, that the squared coefficient of variation of the estimators is effectively constant for a large range of different values of n. This means that many sequencing strategies are very nearly equally efficient for such parameter values. In particular, if only 10 copies of the region were sequenced, the precision would be very close to that for the optimal strategy.

Tables 1 and 2 give details of the optimal strategy and of near-optimal strategies (defined to be those in which the squared coefficient of variation is within 10% of its optimal value) for different total amounts of sequencing effort, defined as the total number, nL, of bases sequenced.

Recall that our analysis requires the assumptions that the underlying population has been panmictic and of constant size throughout the relevant period of its evolution. This evolution is further assumed to be neutral in the region in question, and the population assumed to be in mutation-drift equilibrium. In this setting, for the simple cost function we are considering, the optimal allocation of resources usually requires the sequencing of a small number (typically around five) of large copies of the region in question. Even when this is not true of the optimal strategy, strategies that sequence no more than 10 copies of the region are very close to being optimal.

The discussion above concerns the question of how best to allocate a fixed amount of sequencing effort. A separate issue relates to the appropriate amount of effort to allocate to a particular region. The precision of estimation is an increasing function of the number of bases sequenced, for any fixed sequencing strategy and hence also when comparing the optimal strategies for different total numbers of bases sequenced. There is, however, a trade-off between gains at the region of current interest, and the possibility of examining another locus.

The problem of when to move to a different locus, or, in a study of variability in different regions, of how many loci to examine, does not seem to lend itself to a precise formulation. Different approaches may be appropriate for studies with different goals. In Figure 4 we plot the relative variability of the estimators, under the optimal sequencing strategy, for various different amounts of "total effort". Examination of the figure allows an assessment of the marginal gain from extra sequencing at the region of interest. This can then be compared to the potential additional information that may be obtained by studying a different, unlinked, region of the genome.

We will discuss the actual level of correlation between estimates from distant, but linked, regions of the ge-

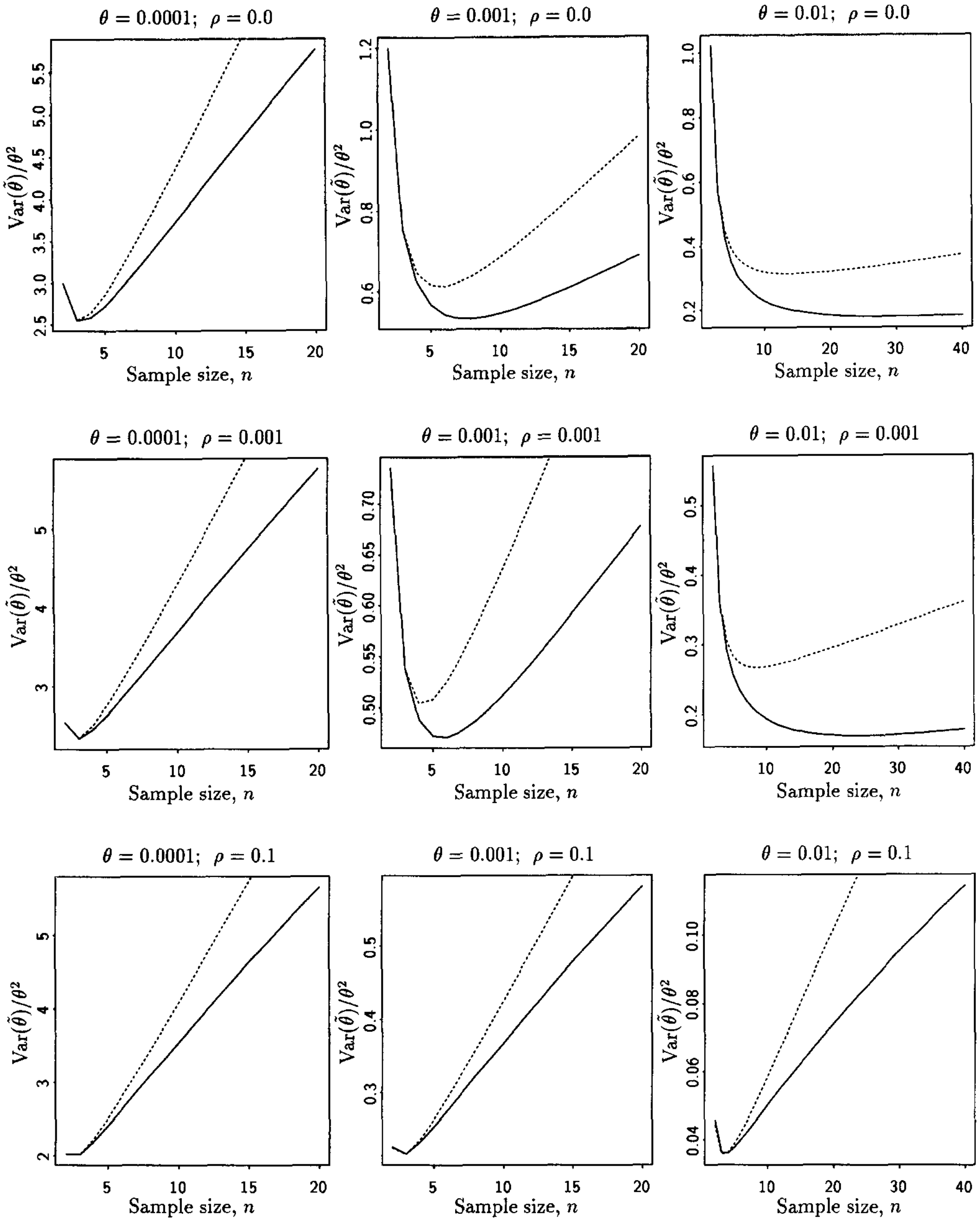

FIGURE 3.—Normalized sampling variance of the estimators $\tilde\theta_W$ (——) and $\tilde\theta_T$ (· · ·) as a function of sample size n, when the total number of sites to be sequenced, nL, is fixed at 10,000.

nome in the next section. For the moment, however, suppose we have an option of studying two regions, A and B say, with the property that the estimates from them are independent. Consider a fixed amount of sequencing effort, say a total of κ bases, and contrast the following alternatives.

1. Having initially sequenced κ bases from region A, sequence κ bases from region B.
2. Having initially sequenced κ bases from region A, sequence an additional κ bases from region A.

Suppose in addition that the underlying evolutionary parameters θ and ρ are the same in each region, so that both strategies use the same amount of effort to measure the "same" level of underlying diversity. Our results, and in particular Figure 4, facilitate a comparison between the two strategies.

Under the first strategy, the variance of either diversity measure is halved. (This is simply the classical result that the variance of the average of two independent and identically distributed quantities is half of the variance of either of the original quantities.) The first strategy will *always* result in a greater reduction in variance than the second because in the second case, for the reasons we have already described, one is effectively averaging positively correlated quantities. It is thus interesting to see how much better off one would be by moving to another region than by improving precision in the current region.

It follows from (4) and (5) for example, that if the total mutation rate $\Theta = L\theta$ of the region initially se-

TABLE 1

Optimal sample size n_{opt} for the Watterson's estimator $\tilde{\theta}_W$ of the scaled mutation rate per site θ (for different total amounts of sequencing effort)

		Total number of sites (nL) to be sequenced											
		2000		4000		6000		10000		20000		50000	
ρ	θ	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range
0.0	0.0001	3	(2, 4)	3	(2, 4)	3	(3, 4)	3	(3, 5)	4	(3, 6)	6	(4, 9)
	0.001	4	(3, 6)	5	(4, 8)	6	(4, 10)	8	(5, 13)	11	(7, 19)	18	(11, 33)
	0.01	10	(7, 19)	16	(9, 29)	20	(12, 37)	28	(15, 52)	43	(23, 85)	80	(40, *)
0.001	0.0001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(3, 4)	3	(3, 5)
	0.001	4	(3, 6)	4	(3, 7)	5	(4, 8)	6	(4, 10)	6	(4, 12)	7	(4, 15)
	0.01	10	(6, 18)	15	(8, 27)	18	(10, 34)	25	(12, 47)	34	(15, *)	55	(18, *)
0.01	0.0001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)
	0.001	3	(3, 4)	3	(3, 5)	3	(3, 5)	3	(3, 5)	3	(3, 5)	3	(3, 5)
	0.01	6	(4, 12)	7	(4, 14)	7	(4, 15)	6	(4, 14)	5	(4, 11)	5	(4, 9)
0.1	0.0001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)
	0.001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)
	0.01	3	(3, 5)	3	(3, 5)	3	(3, 5)	3	(3, 5)	4	(3, 5)	4	(3, 5)

Entries in each cell correspond to the sample size n_{opt} that minimizes the coefficient of variation of the estimator $\tilde{\theta}_W$, and the range of values of n for which the coefficient of variation is within 10% of the optimal value. * signifies that the entry is >100.

quenced (where the sequence length L is chosen to be optimal for a total amount of effort κ) is, say, <0.1, then the effect of strategy 2 will effectively be to halve the variance of the estimate, even in the absence of intragenic recombination. Regarding θ as fixed, this means that if the length of the region initially sequenced is, say, an order of magnitude or more smaller than θ^{-1}, there is little additional gain in precision from moving to the new region.

At the other extreme, if the length L of the region initially sequenced in each sampled individual is very large, we might suppose that the asymptotic rate of decay of the variance of $(\log L)/L$ obtains. In this case the variance will decrease under strategy 2 by a factor of $2(\log L)/(\log L + \log 2) \approx 2$ for large L. Thus, again, both strategies will lead to approximately the same reduction in variance, and there is little additional gain from moving to a new region.

For particular values of κ and the parameters θ and ρ, the strategies may be compared via Figure 4. For example, consider $\theta = 0.001$ and $\rho = 0.001$ with $\kappa = 5000$. The variance of both measures is $\sim 8 \times 10^{-7}$. Under strategy 1, this would be reduced to 4×10^{-7}. On the other hand, if 10,000 bases were sequenced (optimally) from the region, Figure 4 shows that the variance would decrease to somewhat $<5 \times 10^{-7}$. There

TABLE 2

Optimal sample size n_{opt} for the Tajima's estimator $\tilde{\theta}_T$ of the scaled mutation rate per site θ (for different total amounts of sequencing effort)

		Total number of sites (nL) to be sequenced											
		2000		4000		6000		10000		20000		50000	
ρ	θ	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range	n_{opt}	Range
0.0	0.0001	3	(2, 4)	3	(2, 4)	3	(3, 4)	3	(3, 4)	4	(3, 5)	5	(4, 7)
	0.001	4	(3, 5)	4	(3, 6)	5	(4, 7)	6	(4, 9)	7	(5, 12)	10	(6, 20)
	0.01	7	(5, 12)	10	(6, 18)	11	(6, 22)	14	(7, 30)	19	(9, 45)	28	(11, 84)
0.001	0.0001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(3, 4)	3	(3, 4)
	0.001	3	(3, 5)	4	(3, 6)	4	(3, 6)	4	(3, 7)	4	(3, 7)	4	(3, 7)
	0.01	7	(5, 11)	8	(5, 15)	8	(5, 17)	8	(5, 19)	7	(4, 17)	5	(4, 10)
0.01	0.0001	3	(2, 3)	3	(2, 3)	3	(2, 3)	3	(2, 3)	3	(2, 4)	3	(2, 4)
	0.001	3	(3, 4)	3	(3, 4)	3	(3, 4)	3	(3, 4)	3	(3, 4)	3	(3, 4)
	0.01	4	(3, 7)	4	(3, 7)	4	(3, 7)	4	(3, 7)	4	(3, 6)	4	(3, 6)
0.1	0.0001	3	(2, 3)	3	(2, 3)	3	(2, 3)	3	(2, 3)	3	(2, 3)	3	(2, 3)
	0.001	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)	3	(2, 4)
	0.01	3	(3, 4)	3	(3, 4)	3	(3, 5)	3	(3, 5)	3	(3, 5)	4	(3, 5)

Entries in each cell correspond to the sample size n_{opt} that minimizes the coefficient of variation of the estimator $\tilde{\theta}_T$, and the range of values of n for which the coefficient of variation is within 10% of the optimal value.

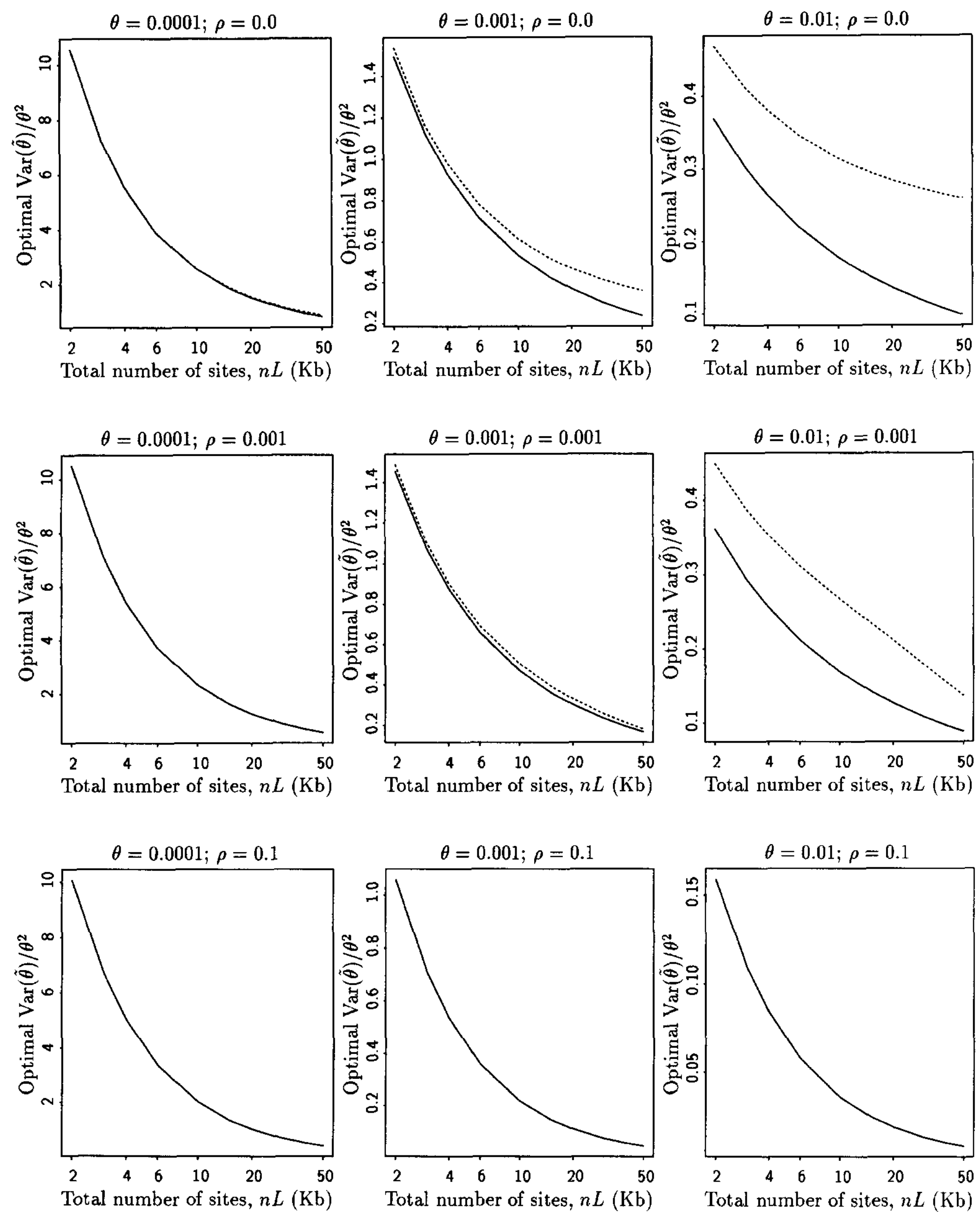

FIGURE 4.—Normalized optimal sampling variance of the estimators $\tilde{\theta}_W$ (——) and $\tilde{\theta}_T$ (· · ·) as a function of the total number of sites to be sequenced, nL (in kb).

is thus some gain from moving to an "independent" region, but rather less than might have been expected. If, instead, 20,000 bases had already been sequenced (κ = 20,000), the variance of the estimators would be ~3 × 10^{-7}. Strategy 1 would reduce this to 1.5 × 10^{-7}, while strategy 2 would only reduce it to ~2 × 10^{-7}. The relative advantage of strategy 1 is thus greater than when 5000 bases had originally been sequenced.

A consequence of our analysis then is that while one will always do better, in terms of increasing the precision of the measures of diversity, by moving to an independent region (with the same values of the underlying parameters) than by extending the current region, this difference is much less marked (for many parameter values) than might have been expected. The question of when to move to a different region is thus likely to turn on other issues. On the one hand, one wants enough information from the current region to give reasonable precision to the estimates of diversity. (Exactly how much precision is appropriate could differ markedly from study to study, depending on the overall goals.) For a given such level of precision, Figure 4 may be used to determine the necessary total amount of effort and Tables 1 and 2, to determine the appropriate sequencing strategy. On the other hand, one reason for examining different regions of the genome is precisely because they may not reflect the same underlying levels of diversity. In this sense, aside from any additional "costs" incurred, there is a substantial benefit to moving to a different region over and above considerations of the precision of estimation.

Another interesting feature of Figure 4 is that under the appropriate optimal strategy, the precisions of the two measures $\tilde{\theta}_W$ and $\tilde{\theta}_T$ are very similar, unless θ is large ($\theta > 0.01$) and ρ is small ($\rho \leq 0.001$). For some parameter values, in fact, they are effectively identical. It is well known that when the intragenic recombination rate is zero, the measure $\tilde{\theta}_W$ is to be preferred to $\tilde{\theta}_T$ in terms of its precision. An important conclusion of our analysis is that in the presence of intragenic recombination, under the optimal sequencing strategy, these two measures are effectively equivalent from the perspective of their sampling variability for most of the parameter values we considered. (Recall that the optimal strategy involves sequencing a few long copies of the region of interest. The equivalence of the precision of the measures is a consequence of this sequencing strategy. As we saw in Figures 1 and 2, the difference in sampling variance between $\tilde{\theta}_W$ and $\tilde{\theta}_T$ is greater for shorter regions and/or large sample sizes.)

CORRELATIONS BETWEEN ESTIMATORS FROM LINKED REGIONS

In this section we assess the correlation between measures of diversity from linked regions of a chromosome. Loosely speaking, the aim is to answer the question of how far apart two regions must be for inferences from them to be approximately independent.

We thus consider two regions of length L, each of which is evolving according to the model described above (and in particular, within each of which there may be recombination). For definiteness we assume that the recombination rate between sites is the same within and between the two regions. If the distance between the regions is D bases, the scaled recombination rate between the two regions is $D\rho$. Figure 5 plots the correlation between the diversity measures from the two regions as a function of the distance, D, between the regions, for the specific case $L = 1000$ and $n = 10$, for various parameter values.

Expressions for the covariance of the diversity measures from different loci are given in the APPENDIX, see (A15) and (A16). The correlation can be calculated from this and the formulae (6) and (7). The value of the correlation depends in a complicated way on each of the parameters L, n, θ and ρ. Nonetheless, the behavior in Figure 5 is typical. In particular, for ρ at least 0.001, estimators from regions 10 kb apart are effectively uncorrelated. For many parameter values rather smaller distances between the loci still lead to effectively uncorrelated estimators.

The correlation depends on the distance between the regions only as a function of the scaled recombination rate between the regions. That is, it is a function only of $D\rho$. Results for a setting in which the recombination rate between the regions, per generation, is r_1, while that within each region is r per generation, can be obtained as the correlation at distance $D = r_1/r$ in Figure 5. (Strictly, the correlation is a function of both $\rho = 4Nr$ and r_1/r.)

KAPLAN and HUDSON (1985, Figure 1) plotted the correlation in tree lengths between two regions, as a function of the scaled recombination rate ($4Nr_1$ in the notation of the previous paragraph) between the regions. They assumed no recombination within the regions, in which case the covariance of tree lengths would (by a simple extension of the argument in the APPENDIX) be proportional to the covariance of the measure $\tilde{\theta}_W$ from the two regions. One novelty in Figure 5 thus relates to the correlation for the measure $\tilde{\theta}_T$. In addition, we have extended the analysis to include intragenic recombination.

DISCUSSION

We consider the possible effects on our conclusions of changes to various of the underlying assumptions.

Throughout, we have assumed that the option is either to sequence a region from the existing sample, which is adjacent to that already sequenced, or to sequence additional copies of the current region from new individuals. In principle, one could aim to get the best of both strategies by sequencing an adjacent region in new individuals. As expected, such an approach does lead to lower variability than either of the separate strategies. On the other hand, the difference between the variability when the region is extended in already sequenced individuals, and that when the adjacent region is sequenced in new individuals, is relatively small (data not shown). Recall that the gain in sequencing new individuals in a particular region is small because of strong positive correlations between the sequences from different individuals. For exactly this reason, the types of the new individuals sequenced in the adjacent region will be strongly positively correlated with the types in that region of those already sequenced. The gain from sequencing new individuals in the new region is thus not much greater than from extending the sequences already obtained.

The particular, simple, assumption about sequencing costs, which underlies the analysis, while not exact, may not be entirely unrealistic. For any other, particular, "cost" function, a similar analysis is in principle straightforward. Our expectation would be that unless the cost function changes markedly, the same broad conclusion, namely that sequencing relatively few, long, copies of the region is appropriate, should still obtain.

Our results are unlikely to remain valid if the demographic assumptions about the population are changed substantially. The effect on the underlying genealogies of such changes, for example variation in population size, and/or geographical structure of the population, is reasonably well understood [see, for example, HUDSON (1991, 1993) or DONNELLY and TAVARÉ (1995)]. In principle, it would thus be possible, at least via simu-

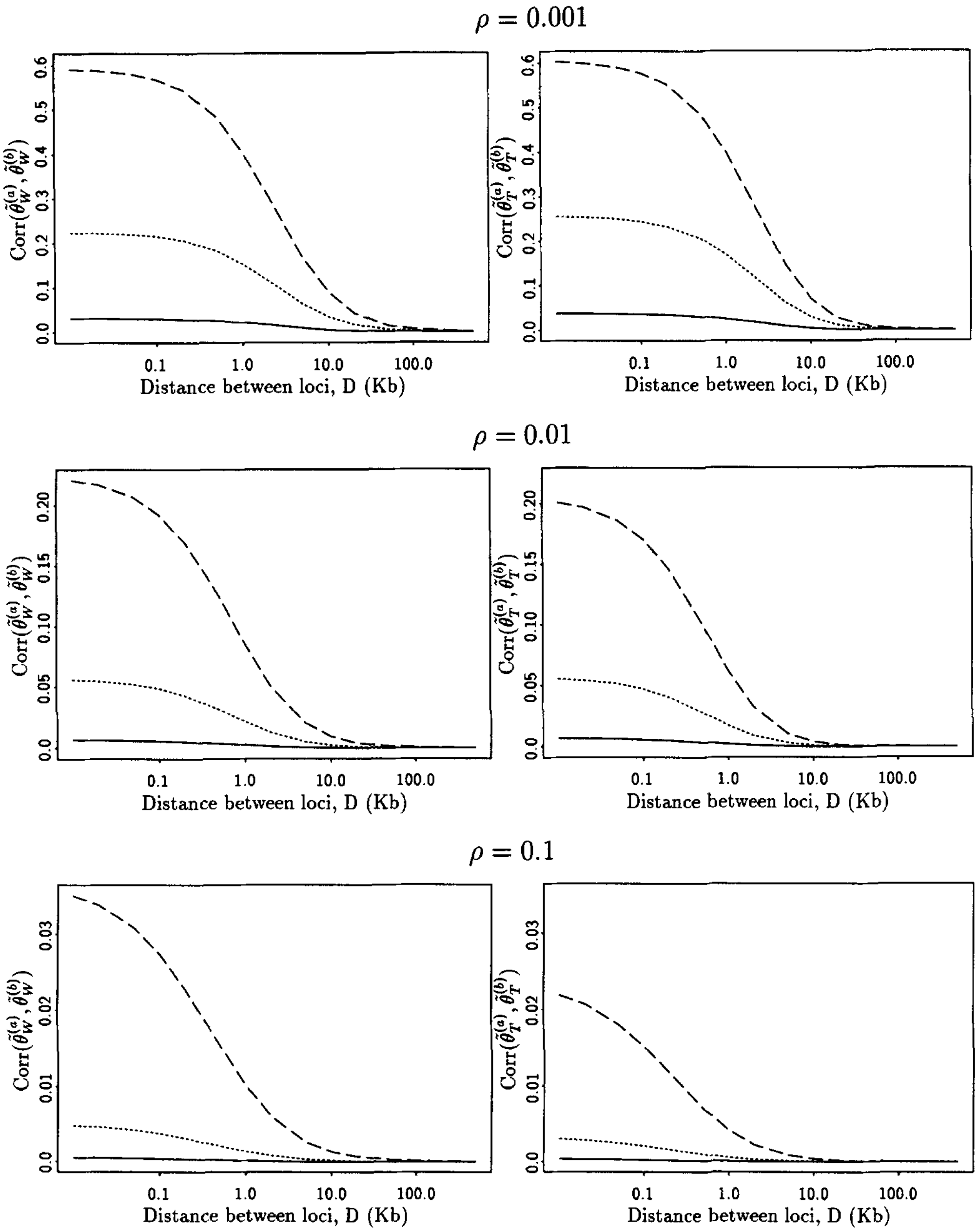

FIGURE 5.—Sampling correlation between the estimators $\tilde\theta_W^{(a)}$, $\tilde\theta_W^{(b)}$ (left column), and $\tilde\theta_T^{(a)}$, $\tilde\theta_T^{(b)}$ (right column) calculated at two distinct loci a and b of the same length $L = 1000$ bp, as a function of distance D between the loci, for a fixed value of the sample size $n = 10$, and values of the scaled recombination rate per site $\rho = 0.001$, 0.01, and 0.1. The three lines correspond to the values of the scaled mutation rate per site $\theta = 0.0001$ (——), $\theta = 0.001$ (· · ·), and $\theta = 0.01$ (- - -).

lation, to extend our results to incorporate specific alternative demographic beliefs. We content ourselves here with a heuristic discussion of the qualitative consequences of some specific assumptions. Note that in general, changes in the underlying assumptions will change the sampling mean of the estimators, so that $\tilde\theta_W$ and $\tilde\theta_T$ will no longer be natural estimators for θ. Exactly how they should be corrected will depend sensitively on the underlying assumptions.

The effect of a population bottleneck, or a rapid expansion (forward in time) of the population, is to change genealogical trees from those predicted by the coalescent to a rather more "star-shaped" topology. In such a setting, sequences in the sample are much more independent than under the coalescent. This reflects the fact that in a star-shaped genealogical tree, distinct sequences share very little of their ancestral history after the common ancestor of the sample. In this case, the decrease in the variability of estimators induced by increasing the sample size is relatively much greater than under the coalescent model. As a consequence, the optimal trade-off between sample size and sequence length will lie much more in the direction of larger samples and smaller sequences than for the coalescent. The quantitative extent of this change will, naturally, depend on the severity of the bottleneck or the amount and rate of population growth. The effect of a selective sweep at a locus closely linked to the region under study is very similar to that of rapid population growth (for example, KAPLAN *et al.* 1989), so that the same conclusion should apply in this context.

In the presence of population subdivision, sequences from different subpopulations, or geographical areas, will tend to be more independent than under the coalescent. Thus, again, there will be more of a gain in sequencing an additional individual than under the coalescent assumptions, provided the individual is taken from a new area. This reinforces the obvious advantages of the strategy of obtaining samples from at least several subpopulations or areas. The intuition behind the analysis above suggests that the sample size within each area need not be too large. The effect of balancing selection at a linked locus is similar to that of some versions of population structure (*e.g.*, Hudson 1993).

Simonsen *et al.* (1995) recently addressed the question of sequencing strategies in the context of testing a neutral hypothesis against that of a selective sweep. Their conclusion was that for Tajima's test, and that particular alternative, "it is better to sequence more individuals than more sites, so long as the number of sites is not too small". The study by Simonsen *et al.* (1995) did not incorporate recombination (either within the sequenced region or between the region and the selected locus). We have seen that the relative advantage of extending the sequenced region over that of sequencing new copies of the region increases with the recombination rate. Their result that sample sizes should be large will thus depend, to some extent, on the lack of recombination. Nonetheless, their result is exactly contrary to our results on the optimal strategy for assessing population diversity from a neutral, panmictic population of constant size. Assessing diversity under neutrality and testing neutrality are two quite different things. There is no *a priori* reason why an experimental design that is good for one purpose should be good for another. Indeed, as discussed above, if there has been a selective sweep, the optimal sequencing strategy for assessing diversity will involve more individuals and fewer sites than in the neutral case, so it is perhaps not surprising that this latter strategy is also most appropriate for testing neutrality against that alternative.

CONCLUSIONS

We have considered the precision of two commonly used measures of genetic diversity at the DNA sequence level, as a function of both the sample size and the length of the region sequenced. The first of these measures, $\tilde{\theta}_W$, defined via (1) and (3), is based on the number of segregating sites in the sample. The second, $\tilde{\theta}_T$, defined via (2) and (3), is the average number of pairwise differences in the sample. Our analysis applies in the context of a large panmictic population that has been of constant size throughout its evolution. Evolution in the region of interest is assumed to be neutral. Recombination is allowed, at constant rate, between any pair of adjacent sites in the region.

It is well known that for sequences of fixed length, L, there is relatively little gain in precision as the sample size, n, is increased. The sampling variance of $\tilde{\theta}_W$ does converge to zero as $n \to \infty$, but extremely slowly, while that of $\tilde{\theta}_T$ converges to a nonzero constant as $n \to \infty$. This results from the fact that sampled sequences are highly positively correlated, because they share much of their ancestral history.

There is, however, another natural "direction" in which to increase the available information: the length of the region sequenced may be increased while keeping the sample size fixed. In this setting, provided the scaled recombination rate $\rho > 0$, the asymptotic behavior is much more encouraging. The variance of both estimators converges to zero at least as fast as $(\log L)/L$. The intuition is that in the presence of recombination, parts of the evolution at different sites are independent. Thus, increasing the length of sequence corresponds loosely to gaining independent replications of the underlying evolutionary process.

We considered the optimal allocation of sequencing resources when the number of bases to be sequenced is held fixed, and the cost of sequencing an additional base is the same regardless of whether it arises from a new individual or from the extension of an existing sequence. In this context, for the evolutionary models we are considering, the optimal strategy depends somewhat on the underlying values of the mutation and recombination parameters θ and ρ. Nonetheless, unless θ is large (say 0.01), the optimal sampling scheme involves sequencing very few (typically around five) long copies of the region of interest. Even for large θ, strategies that involve sample sizes of, say, 10, are close to being optimal.

Having decided how best to allocate a predetermined amount of sequencing effort to a particular chromosomal region, there is a separate question of the appropriate amount of effort to devote to that region. While more effort will increase the amount of information available from the region, this will be at the cost of information that could be obtained by sequencing a different locus. The trade-off here is likely to depend on the goals of the study. For a large range of parameter values, the reduction in the variability of the estimators when a fixed additional amount of sequencing is done in the region under study is surprisingly close to that which would be obtained were the additional effort directed to a distinct region.

One conclusion of our study is that unless θ is large ($\theta > 0.01$) and ρ is small ($\rho \leq 0.001$), the precision of the estimators $\tilde{\theta}_W$ and $\tilde{\theta}_T$ is very similar under the optimal sequencing strategy. (The optimal strategy may not be the same for each estimator.) Thus, for most of the parameter values considered here, there are not strong reasons to prefer one measure to the other on the grounds of efficiency, provided an optimal (or close to optimal) sequencing strategy is adopted. This conclusion is related to the fact that optimal sequencing strategies typically involve small samples of long sequences. The conclusion is false, and $\tilde{\theta}_W$ preferable in terms of precision, for many other sampling schemes.

The extent to which estimates from two regions, or loci, are uncorrelated is, of course, a function of the genetic distance between the regions, and of the effective population size. Under the assumption of a constant recombination rate between sites, which may be plausible over some physical distances but not on longer scales, the correlation is a function of the number of base pairs, D, between the regions and the scaled recombination rate between sites ρ. Loosely speaking, the correlation is small, and the estimates effectively uncorrelated, if the product $D\rho$ is at least 10.

Taken together, our results may be helpful in designing sequencing studies aimed at assessing molecular genetic diversity. If the desired level of precision for estimating diversity in a particular region is specified, then for particular parameter values, the required total number of bases to be sequenced may be obtained from Figure 4. The appropriate sequencing strategy can then be deduced from Tables 1 and 2. The extent of independence between different regions is given in Figure 5.

Exactly which strategies are optimal, and the associated variability, depends on the underlying evolutionary parameters θ and ρ. In practice these may not be known in advance. One general conclusion, valid for a large range of parameter values, is that strategies that involve sequencing relatively few (five to 10) long copies of the region are to be preferred to those with larger sample sizes and correspondingly smaller sequence length. At a finer level, the dependence of optimal strategies and associated precision on the underlying parameters is very smooth, so that "good" strategies can be chosen if the parameters are only known up to, say, an order of magnitude.

Our conclusions *are* likely to be sensitive to gross violations of the underlying assumptions of panmixia and neutrality. In the presence of population bottlenecks, recent population expansions, or selective sweeps at linked loci, optimal strategies will tend much more toward large sample sizes and relatively smaller sequence length. SIMONSEN *et al.* (1995) note that the same conclusion is valid if one wishes to use Tajima's test to detect departures from neutrality in the direction of a selective sweep. If there is underlying population structure, then there are obvious advantages to sampling from as many different subpopulations as is practicable. An important caveat is thus that the best sequencing strategy will in general depend on the primary goal of an experiment, and what is known, or believed, about the evolutionary or demographic processes that have generated the current diversity. Strategies that are good for one purpose, or in one context, may not be good for another purpose or in another setting.

The first author was supported in part by Public Health Service grant 5 R01 GM46800 and in part by National Institutes of Health grant HL49596. The second author was supported in part by National Science Foundation (NSF) grant DMS-9505129 and in part by The University of Chicago Block Fund. Computations for this paper were performed using computer facilities supported in part by NSF grant DMS 89-05292 awarded to the Department of Statistics at The University of Chicago, and by The University of Chicago Block Fund.

LITERATURE CITED

CHAKRABORTY, R., and R. C. GRIFFITHS, 1982 Correlation of heterozygosity and the number of alleles in different frequency classes. Theor. Popul. Biol. **21:** 205–218.

DONNELLY, P., and S. TAVARÉ, 1995 Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. **29:** 401–421.

ETHIER, S. N., and R. C. GRIFFITHS, 1990 On the two-locus sampling distribution. J. Math. Biol. **29:** 131–159.

FU, Y. X., 1994 Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequencies. Genetics **138:** 1375–1386.

GRIFFITHS, R. C., 1981 Neutral two-locus multiple allele models with recombination. Theor. Popul. Biol. **19:** 169–186.

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

HUDSON, R. R., 1991 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

HUDSON, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution: Introduction to Molecular Paleopopulation Biology*, edited by N. TAKAHATA and A. G. CLARK. Sinauer Associates, Inc., Sunderland, MA.

KAPLAN, N. L., and R. R. HUDSON, 1985 The use of sample genealogies for studying a selectively neutral *m*-loci model with recombination. Theor. Popul. Biol. **28:** 382–396.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

SIMONSEN, K. L., G. A. CHURCHILL and C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: R. R. HUDSON

APPENDIX

Variances of the estimators with intragenic recombination: We derive the variances of the estimators $\tilde{\theta}_W$ and $\tilde{\theta}_T$ using coalescent theory. For background on the coalescent, including the results used below, see for example HUDSON (1991, 1993), or DONNELLY and TAVARÉ (1995). Consider first $\tilde{\theta}_W$. Recall that S_n is the number of segregating sites in the n sequences of length L. By (1) and (3),

$$\mathrm{Var}(\tilde{\theta}_W) = \frac{\mathrm{Var}(S_n)}{L^2(\sum_{i=1}^{n-1} i^{-1})^2}, \tag{A1}$$

so we consider $\mathrm{Var}(S_n)$.

Consider a particular site, k say, in the sequence. There will be a genealogical tree that describes the ancestral history, back to their common ancestor, of the n copies of that site in the sample. Denote the total length of this tree by $T_n^{(k)}$, and the number of mutation events on the tree by $M_n^{(k)}$. For any k, $T_n^{(k)}$ are identically distributed (though *not* independent), with

$$E(T_n^{(k)}) = 2\sum_{i=1}^{n-1} i^{-1} \qquad \text{(A2)}$$

and

$$\mathrm{Var}(T_n^{(k)}) = 4\sum_{i=1}^{n-1} i^{-2}. \qquad \text{(A3)}$$

Conditional on the respective tree lengths, $T_n^{(k)}$ and $T_n^{(j)}$, the numbers of mutations $M_n^{(k)}$ and $M_n^{(j)}$ at different sites k and j are independent.

By the infinite sites assumption, S_n is equal to the total number of mutations on the genealogical trees for each site:

$$S_n = \sum_{k=1}^{L} M_n^{(k)}.$$

Thus

$$\mathrm{Var}(S_n) = \sum_{k=1}^{L} \mathrm{Var}(M_n^{(k)}) + 2\sum_{k=1}^{L}\sum_{j=k+1}^{L} \mathrm{Cov}(M_n^{(k)}, M_n^{(j)}). \qquad \text{(A4)}$$

Observe that, under the infinite sites assumption, $M_n^{(k)}$ takes values of 1 or 0 only, depending on whether or not there was a mutation at the kth site in the ancestral history of the sample. Then,

$$P(M_n^{(k)} = 0 \mid T_n^{(k)}) = e^{-\theta T_n^{(k)}/2},$$

and, hence,

$$\begin{aligned}\mathrm{Var}(M_n^{(k)}) &= P(M_n^{(k)} = 0)(1 - P(M_n^{(k)} = 0)) \\ &= E(e^{-\theta T_n^{(k)}/2})(1 - E(e^{-\theta T_n^{(k)}/2})).\end{aligned}$$

An approximation for $E(e^{-\theta T_n^{(k)}/2})$ follows immediately from (9) and (10):

$$\begin{aligned}E(e^{-\theta T_n^{(k)}/2}) &\approx E(1 - \theta T_n^{(k)}/2 + \theta^2 T_n^{(k)^2}/8) \approx 1 \\ &- \theta \sum_{i=1}^{n-1} i^{-1} + \frac{1}{2}\theta^2\left(\left(\sum_{i=1}^{n-1} i^{-1}\right)^2 + \sum_{i=1}^{n-1} i^{-2}\right).\end{aligned}$$

Thus, for small θ,

$$\begin{aligned}\mathrm{Var}(M_n^{(k)}) &\approx \theta \sum_{i=1}^{n-1} i^{-1} \\ &- \frac{1}{2}\theta^2\left(3\left(\sum_{i=1}^{n-1} i^{-1}\right)^2 + \sum_{i=1}^{n-1} i^{-2}\right) \approx \theta \sum_{i=1}^{n-1} i^{-1}. \qquad \text{(A5)}\end{aligned}$$

Further,

$$\begin{aligned}\mathrm{Cov}(M_n^{(k)}, M_n^{(j)}) &= E(\mathrm{Cov}(M_n^{(k)}, M_n^{(j)} \mid T_n^{(k)}, T_n^{(j)})) \\ &+ \mathrm{Cov}(E(M_n^{(k)} \mid T_n^{(k)}, T_n^{(j)}), E(M_n^{(j)} \mid T_n^{(k)}, T_n^{(j)})) \\ &= 0 + \mathrm{Cov}((1 - e^{-\theta T_n^{(k)}/2}), (1 - e^{-\theta T_n^{(j)}/2})) \\ &\approx \frac{\theta^2}{4}\mathrm{Cov}(T_n^{(k)}, T_n^{(j)}), \qquad \text{(A6)}\end{aligned}$$

in view of the conditional independence of $M_n^{(k)}$ and $M_n^{(j)}$ given $T_n^{(k)}$ and $T_n^{(j)}$, and θ being small.

The covariance of $T_n^{(k)}$ and $T_n^{(j)}$ is exactly the covariance of the lengths of the genealogical trees in a two-locus model in which the recombination rate between the loci is $(j - k)\rho$. (Here and below, our discussion of two locus models applies to the case in which there is no recombination within either locus.) Define $F(0, 0, c; z)$ to be the covariance between the lengths of the genealogical trees at each locus in a two-locus model in which the same c chromosomes are sampled at each locus, and the scaled recombination rate between the loci is z. More generally, define $F(a, b, c; z)$ to be the covariance between the lengths of the genealogical trees at each locus in a two locus model in which $a + c$ chromosomes are sampled at the first locus and $b + c$ chromosomes are sampled at the second locus in such a way that exactly c chromosomes are common to both samples, and the scaled recombination rate between the loci is z. The total sample size is thus $n = a + b + c$.

It can be shown that for $0 \le z < \infty$, the function $F(a, b, c; z)$ satisfies the linear system

$$\begin{aligned}F(a, b, c; z) = \frac{1}{\beta_n}[&r_1 F(a + 1, b + 1, c - 1; z) \\ &+ r_2 F(a - 1, b - 1, c + 1; z) + r_3 F(a - 1, b, c; z) \\ &+ r_4 F(a, b - 1, c; z) + r_5 F(a, b, c - 1; z) + R_n],\end{aligned} \qquad \text{(A7)}$$

where $n = a + b + c$, $\beta_n = (n(n - 1) + cz)/2$, $r_1 = cz/2$, $r_2 = ab$, $r_3 = ac + a(a - 1)/2$, $r_4 = bc + b(b - 1)/2$, $r_5 = c(c - 1)/2$, and $R_n = 2c(c - 1)/((a + c - 1)(b + c - 1))$. The initial conditions are $F(a, b, c; z) = 0$ whenever $a < 0$, or $b < 0$, or $c < 0$, or $a + c < 2$, or $b + c < 2$.

The recursion (A7) is derived by conditioning on the first event in the genealogical history of the sample. Such an event is either a recombination or a coalescence. In the latter case there are four possibilities: a coalescence of two of the chromosomes sampled only at the first locus, a coalescence of two of the chromosomes sampled only at the second locus, a coalescence of two of the chromosomes sampled at both loci, or a coalescence of one of the chromosomes sampled only at the first locus with one of the chromosomes sampled only at the second locus. Kaplan and Hudson (1985) used exactly this technique to derive a recursion for the expected value of the product of the tree lengths at two linked loci. We refer the reader to that paper for details of the method. The recursion (A7) can be shown to be equivalent to the one derived earlier by Kaplan and Hudson (1985) (except for some misprints there). It appears somewhat easier to implement in practice. The recursion is three-dimensional. We adopted the method described in Ethier and Griffiths (1990) for its numerical solution.

Write $F_n(z)$ for $F(0, 0, n; z)$. It follows from (A6) and the definition of F_n, that the second term, Σ_2 say, in (A4) is

$$\Sigma_2 = \frac{\theta^2}{2}\sum_{k=1}^{L}\sum_{j=k+1}^{L} F_n((j-k)\rho) = \frac{\theta^2}{2}\sum_{m=1}^{L-1}(L-m)F_n(m\rho)$$

$$\approx \frac{L^2\theta^2}{2}\int_0^1 (1-x)F_n(L\rho x)\,dx, \quad \text{(A8)}$$

for large L. The result (6) then follows from (8), (A4), (A5), and (A8).

The integral approximation (A8) is identical (up to a linear change of variables) to the one derived in HUDSON (1983) (see also KAPLAN and HUDSON (1985) for further details) for a slightly different model of an infinite sites locus with recombination than the one adopted in this paper. We have modeled the locus as comprising a large number of linked sites, with recombination taking place between any two adjacent sites, and (according to the infinite sites assumption) no more than one mutation per site. In contrast, Hudson's model considers each site *itself* to be an infinite sites sublocus, *i.e.*, an infinite sequence of completely linked sites. Then the L consecutively arranged subloci together constitute the locus under consideration, recombination taking place only between subloci. The limiting properties of Hudson's model, including the approximation (A8), are obtained by letting $L \to \infty$ while holding the total mutation rate Θ and the total recombination rate R fixed. Up to the level of approximation we have used to derive the expression (6) (ignoring terms of order θ^2/L and higher), the two approaches yield identical results. We note, however, that they differ in higher order terms.

Next consider $\mathrm{Var}(\tilde{\theta}_T)$. Write

$$\tilde{\theta}_T = \frac{2}{Ln(n-1)}\sum_{l=1}^{L}\sum_{i<j}\delta_{ij}^{(l)}, \quad \text{(A9)}$$

where $\delta_{ij}^{(l)}$ equals 1 if the ith and jth sequences in the sample differ at the lth site, and 0 otherwise. Under the infinite sites assumption, $\delta_{ij}^{(l)}$ will equal 0 if and only if there has been no mutation at the lth site in the ancestry of sequences i and j at that site since their common ancestor. Write $T_{ij}^{(l)}$ for twice the (coalescent) time since the common ancestor of these sequences at this site, and note that this is also the total length of the genealogical tree for this sample of two sites. Then,

$$P(\delta_{ij}^{(l)} = 0 \mid T_{ij}^{(l)}) = e^{-\theta T_{ij}^{(l)}/2}.$$

Thus

$$\mathrm{Var}(\tilde{\theta}_T) = L^{-2}\sum_{l=1}^{L}\mathrm{Var}\left(\binom{n}{2}^{-1}\sum_{i<j}\delta_{ij}^{(l)}\right) + \left(\frac{2}{Ln(n-1)}\right)^2$$

$$\times\, 2\sum_{l<m}\sum_{i<j}\sum_{h<k}\mathrm{Cov}(\delta_{ij}^{(l)}, \delta_{hk}^{(m)}). \quad \text{(A10)}$$

On reflection, it is apparent that the first term in (A10) is (L^{-1} times) the variance of the sample heterozygosity in an infinite alleles model with mutation rate θ. The first term is thus (*e.g.*, CHAKRABORTY and GRIFFITHS 1982)

$$\frac{1}{L}\left[\frac{1}{n(n-1)}\left(\frac{2}{1+\theta} + \frac{8(n-2)}{(1+\theta)(2+\theta)}\right.\right.$$

$$\left.\left.+\frac{(n-2)(n-3)(6+\theta)}{(1+\theta)(2+\theta)(3+\theta)}\right) - \frac{1}{(1+\theta)^2}\right]$$

$$\approx \theta\,\frac{n+1}{3L(n-1)}, \quad \text{(A11)}$$

for small θ.

Next,

$$\mathrm{Cov}(\delta_{ij}^{(l)}, \delta_{hk}^{(m)}) = E(\mathrm{Cov}(\delta_{ij}^{(l)}, \delta_{hk}^{(m)} \mid T_{ij}^{(l)}, T_{hk}^{(m)}))$$

$$+\, \mathrm{Cov}(E(\delta_{ij}^{(l)} \mid T_{ij}^{(l)}, T_{hk}^{(m)}), E(\delta_{hk}^{(m)} \mid T_{ij}^{(l)}, T_{hk}^{(m)}))$$

$$= 0 + \mathrm{Cov}((1 - e^{-\theta T_{ij}^{(l)}/2}), (1 - e^{-\theta T_{hk}^{(m)}/2}))$$

$$\approx \frac{\theta^2}{4}\mathrm{Cov}(T_{ij}^{(l)}, T_{hk}^{(m)}), \quad \text{(A12)}$$

since θ is small. The covariance in (A12) relates to the tree length for samples of size $n = 2$ individuals from two loci between which the recombination rate is $(m - l)\rho$. Thus, recalling the definition of $F(a, b, c; z)$,

$$\mathrm{Cov}(T_{ij}^{(l)}, T_{hk}^{(m)})$$

$$= \begin{cases} F(0, 0, 2; (m-l)\rho) & \text{if } \{i, j\} = \{h, k\}, \\ F(1, 1, 1; (m-l)\rho) & \text{if } \{i, j\} \text{ and } \{h, k\} \text{ have one element in common}, \\ F(2, 2, 0; (m-l)\rho) & \text{if } \{i, j\} \text{ and } \{h, k\} \text{ are distinct}. \end{cases}$$

For samples of size two at each locus, the recursion (A7) reduces to a linear system of three equations. The solution is

$$F(0, 0, 2; z) = \frac{4(z+18)}{z^2 + 13z + 18},$$

$$F(1, 1, 1; z) = \frac{24}{z^2 + 13z + 18},$$

$$F(2, 2, 0; z) = \frac{16}{z^2 + 13z + 18}. \quad \text{(A13)}$$

The formula (A13) for the covariance for a sample of size two from a two locus model is originally due to GRIFFITHS (1981).

Substituting these exact expressions for the covariances for samples of size two into (A12), and then into (A10), collecting terms and approximating the resulting sum by an integral as in the derivation of $\mathrm{Var}(\tilde{\theta}_W)$, and also substituting (A11) into (A10), gives

$$\mathrm{Var}(\tilde{\theta}_T) \approx \frac{\theta(n+1)}{3L(n-1)} + \frac{4\theta^2}{n(n-1)} \times \int_0^1 (1-x)\,\frac{L\rho x + C}{(L\rho x)^2 + 13L\rho x + 18}\,dx, \quad \text{(A14)}$$

where $C = 2(n^2 + n + 3)$. The approximation applies to large L and small θ. The formula (7) follows on evaluating the integral.

Similar arguments can be used to approximate the covariance between estimates from linked loci. Write $\tilde{\theta}_W^{(a)}$ and $\tilde{\theta}_W^{(b)}$ for the diversity measures based on segregating sites at two loci a and b, and analogously for $\tilde{\theta}_T^{(a)}$ and $\tilde{\theta}_T^{(b)}$. For convenience we assume that the length L of the region sequenced, and the sample size n, is the same at each locus, but the generalization is straightforward. Then

$$\mathrm{Cov}(\tilde{\theta}_W^{(a)}, \tilde{\theta}_W^{(b)}) \approx \frac{\theta^2}{4(\Sigma_{i=1}^{n-1} i^{-1})^2} \times \int_0^2 (1 - |1-x|)F_n(D\rho + L\rho x)\,dx, \quad \text{(A15)}$$

and

$$\begin{aligned}\mathrm{Cov}(\tilde{\theta}_T^{(a)}, \tilde{\theta}_T^{(b)}) &\approx \frac{2\theta^2}{n(n-1)} \int_0^2 (1-|1-x|) \times \frac{D\rho + L\rho x + C}{(D\rho + L\rho x)^2 + 13(D\rho + L\rho x) + 18}\,dx \\ &= \frac{\theta^2}{n(n-1)(L\rho)^2}\,[(D_2\rho(2C-13) + 13C - 133) \\ &\quad \times G_1(D_2\rho, D_1\rho) - (D\rho(2C-13) + 13C - 133) \\ &\quad \times G_1(D_1\rho, D\rho) + (D_2\rho - C + 13)G_2(D_2\rho, D_1\rho) \\ &\quad - (D\rho - C + 13)G_2(D_1\rho, D\rho)], \quad \text{(A16)}\end{aligned}$$

where $D_1 = D + L$, $D_2 = D + 2L$,

$$G_1(x, y) = \frac{1}{9.85}\log\left(\frac{(x+1.58)(y+11.42)}{(x+11.42)(y+1.58)}\right),$$

and

$$G_2(x, y) = \log\left(\frac{x^2 + 13x + 18}{y^2 + 13y + 18}\right).$$

We note in passing that the preceeding method may be used to approximate the covariance in the sample heterozygosities at two linked loci. Write $H^{(l)}$ and $H^{(m)}$ for the sample heterozygosities at two loci, labeled l and m, between which the recombination rate is γ per generation. We assume there is no intragenic recombination within either locus. Define $\delta_{ij}^{(l)}$ to be 1 if the ith and jth sequences differ at locus l, and analogously for $\delta_{hk}^{(m)}$. An analysis analogous to the one above gives

$$\begin{aligned}\mathrm{Cov}(H^{(l)}, H^{(m)}) &= 2\binom{n}{2}^{-2} \sum_{i<j}\sum_{h<k} \mathrm{Cov}(\delta_{ij}^{(l)}, \delta_{hk}^{(m)}) \\ &= 2\binom{n}{2}^{-1} \theta^2 \frac{\Gamma + 2(n^2 + n + 3)}{\Gamma^2 + 13\Gamma + 18},\end{aligned}$$

in which $\Gamma = 4N\gamma$, and θ is now the scaled *total* mutation rate at each locus, assumed to be the same for both loci.

Asymptotics as sequence length increases: We now suppose θ and ρ are fixed, with $\rho > 0$, and examine the behavior of $\mathrm{Var}(\tilde{\theta}_W)$ and $\mathrm{Var}(\tilde{\theta}_T)$ as the sequence length L increases with the sample size n held fixed. We assume that the infinite sites assumption remains valid as L increases.

Consider first the pairwise difference estimator $\tilde{\theta}_T$. It is evident that the first term in (7), as well as the second and third terms in the square brackets in (7), are of order L^{-1}. Inspection also shows that the first term in the square brackets in (7) is of order $(\log L)/L$.

Now consider $\mathrm{Var}(\tilde{\theta}_W)$, and suppose initially that $n = 2$. In this case, $\tilde{\theta}_W$ and $\tilde{\theta}_T$ coincide, so by the argument in the preceeding paragraph, $\mathrm{Var}(\tilde{\theta}_W)$ decays to zero at a rate of $(\log L)/L$. It is intuitively clear that for given θ, ρ, and L, $\mathrm{Var}(\tilde{\theta}_W)$ is greatest for samples of size $n = 2$. (This is also evident from the plots in Figures 1 and 2.) In this case, the rate of decrease of $\mathrm{Var}(\tilde{\theta}_W)$ must be at least $(\log L)/L$ for any value of the sample size n.