Two-Locus Sampling Distributions and Their Application

Richard R. Hudson

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received March 1, 2001
Accepted for publication August 31, 2001

ABSTRACT

Methods of estimating two-locus sample probabilities under a neutral model are extended in several ways. Estimation of sample probabilities is described when the ancestral or derived status of each allele is specified. In addition, probabilities for two-locus diploid samples are provided. A method for using these two-locus probabilities to test whether an observed level of linkage disequilibrium is unusually large or small is described. In addition, properties of a maximum-likelihood estimator of the recombination parameter based on independent linked pairs of sites are obtained. A composite-likelihood estimator, for more than two linked sites, is also examined and found to work as well, or better, than other available *ad hoc* estimators. Linkage disequilibrium in the Xq28 and Xq25 region of humans is analyzed in a sample of Europeans (CEPH). The estimated recombination parameter is about five times smaller than one would expect under an equilibrium neutral model.

LINKAGE disequilibrium is widely recognized as an important aspect of variation in natural populations (Lewontin 1964, 1974; Langley *et al.* 1974; Langley 1977). Despite this recognition, there appears to be no consensus about how to analyze linkage disequilibrium or even to summarize the levels of observed linkage disequilibrium when two or more polymorphic sites are observed in a sample of chromosomes. One approach has been to calculate D^2 or r^2 for all pairs of polymorphic sites and plot these values as a function of the distance between each pair of sites. (For examples, see Langley 1977; Chakravarti *et al.* 1984; Langley *et al.* 2000; Taillon-Miller *et al.* 2000.) Since the moments of these summary statistics are known at least approximately under standard neutral models (Ohta and Kimura 1969, 1971; Kimura and Ohta 1971; Hill 1975), this has been useful. However, much information is lost in these summary statistics. An alternative analysis consists of reporting the P value of an exact test of independence for all pairs of sites (*e.g.*, Macpherson *et al.* 1990; Langley *et al.* 2000; Vieira and Charlesworth 2000; however, see Lewontin 1995 for an alternative to this approach). Unfortunately, this approach gives little sense of whether observed levels of linkage disequilibrium are higher or lower than expected for pairs of tightly linked sites.

Recently, methods for estimating likelihoods under simple neutral models have been introduced (Griffiths and Marjoram 1996; Kuhner *et al.* 2000; Nielsen 2000). In principle these methods should allow the most powerful analyses to be carried out on samples with multiple linked polymorphic sites. However, at present, these Monte Carlo methods are extremely computationally intensive, and it has been difficult to assess when a valid estimate of the likelihood is obtained and even more difficult to assess the properties of any statistical inference based on these methods.

In summary, quantifying and interpreting observed patterns of linkage disequilibrium remain a challenge. To address this challenge, we propose in this article that one consider polymorphic sites in pairs and utilize likelihood methods appropriate for analyzing a pair of polymorphic sites. That is, we suggest that it may be of use to interpret observed two-site sample configurations in light of the two-site sampling distribution under a simple neutral model, without summarizing the data in a statistic such as D^2 or r^2. When more than two linked polymorphisms appear in a data set, this approach will entail some loss of information, but a great deal is gained in tractability relative to the full multisite-likelihood approach. In this article, we describe some methods of calculating (or estimating) two-site sampling distributions and some applications of these distributions for the analysis of samples from natural populations.

Although methods of calculating or estimating two-locus sample probabilities have been previously described (Golding 1984; Hudson 1985; Ethier and Griffiths 1990), very little use has been made of these distributions, in part because of the computational effort that has been necessary to obtain these probabilities. However, even inexpensive desktop computers are now sufficiently fast to calculate these probabilities, at least for small sample sizes. Furthermore, the required sampling distributions can be made available over the Internet.

In addition to describing existing methodology, we extend the methods in several ways. These include con-

Address for correspondence: Department of Ecology and Evolution, 1101 E. 57th St., University of Chicago, Chicago, IL 60637. E-mail: rr-hudson@uchicago.edu

sidering samples in which the ancestral/derived states of alleles are taken into account. Also, two-locus diploid sample probabilities are calculated. In addition, we describe how these distributions can be used to assess observed levels of linkage disequilibrium between sites and how they may be used to estimate the recombination rate parameter of a neutral model.

THE MODEL AND NOTATION

We consider a selectively neutral two-locus random union of gametes model with discrete generations and Wright-Fisher sampling to produce succeeding generations (Karlin and McGregor 1968; Ewens 1979; Griffiths 1981). The population size, assumed constant, is denoted N. We assume that each locus has the same neutral mutation rate, although relaxing this assumption is trivial. An infinite-allele model of mutation is assumed, although we focus primarily on the case in which mutation rates are small, in which case the model becomes essentially the same as the infinite-sites mutation model. The neutral mutation rate at each locus is denoted u, and the recombination rate between the two loci is denoted r. For large populations the sampling properties are functions of the composite parameters, $4Nu$ ($\equiv \theta$) and $4Nr$ ($\equiv \rho$) (Ohta and Kimura 1969, 1971; Hill 1975).

We focus our attention on samples with exactly two alleles at each locus. The two alleles at the first locus are designated A_0 and A_1; the two alleles at the other locus are designated B_0 and B_1. (At this point the labeling is arbitrary, although later, when ancestral and mutant alleles are specified, the labeling will be meaningful.) A sample of n gametes is randomly drawn from a population at stationarity under the neutral model. The unordered sample configuration is denoted by $\mathbf{n} = (n_{00}, n_{01}, n_{10}, n_{11})$, where n_{ij} is the number of sampled gametes that carry allele A_i at the A locus and allele B_j at the B locus. Hence, $n_{00} + n_{01} + n_{10} + n_{11} = n$. The frequencies of the A_1 allele and the B_1 allele in the sample are $p_1 = (n_{10} + n_{11})/n$ and $q_1 = (n_{01} + n_{11})/n$, respectively, and the frequency in the sample of the A_1B_1 gamete is $p_{11} = n_{11}/n$. In this notation, $D = p_{11} - p_1q_1$ and $r^2 = D^2/(p_1(1 - p_1)q_1(1 - q_1))$ are two commonly employed sample measures of linkage disequilibrium.

The probability of a particular sample configuration, $\mathbf{n} = (i, j, k, l)$, is denoted $q_u((i, j, k, l); \theta, \rho)$ or when no ambiguity results as $q_u(\mathbf{n}; \theta, \rho)$. This sample probability corresponds to the probability given by Ethier and Griffiths (1990, Equation 2.14) and the quantity ΦM of Golding (1984). We note that $q_u((i, j, k, l); \theta, \rho) = q_u((i, k, j, l); \theta, \rho)$, since we assume that the mutation rate is the same at the two loci. It is $q_u(\mathbf{n}; \theta, \rho)$ and closely related probabilities that are the main foci of this article. Because we are interested in polymorphism at single nucleotide sites, the case of very small θ is of primary interest, and most results will be for the limiting case as $\theta \rightarrow 0$.

OBTAINING SAMPLE PROBABILITIES

Recursion equations method: Numerical values of $q_u(\mathbf{n}; \theta, \rho)$ can be obtained for small samples by solving a recursion, originally due to Golding (1984) and further analyzed by Ethier and Griffiths (1990). The reader should consult these articles for details. For sample sizes >40, the linear systems of equations that need to be solved become quite large. For example, with a sample of size 40, the last set of equations that must be solved has >20,000 equations. However, the system is sparse, having only nine or fewer nonzero coefficients in each equation. A program to numerically solve Golding's recursion was written by the author and is available at http://home.uchicago.edu/~rhudson1. The program utilizes a conjugate gradient method and indexed storages of the sparse matrices as described by Press *et al.* (1992) to solve the linear systems.

Random-genealogies Monte Carlo method: An alternative to solving Golding's recursion is to estimate the two-locus sample probabilities by the method of Hudson (1985). This method is practical for samples of sizes up to 100 and perhaps somewhat larger. Briefly, the estimate is obtained by generating a large number of independent two-locus genealogies (under the neutral model with the appropriate value of ρ) using standard coalescent machinery (Hudson 1983). For each genealogy, one calculates the probability of the sample configuration of interest. The average of these probabilities is an estimate of $q_u(\mathbf{n}; \theta, \rho)$. Because it is of use later, we describe the method in more detail. Before proceeding with this description we note that Monte Carlo Markov chain methods, such as that of Nielsen (2000), are likely to be very much faster than the method described below. Nielsen's method estimates essentially the same probability that we consider here but can be used on the much more difficult problem of more than two linked sites. However, for estimating the probabilities of all possible configurations for a pair of sites and a given sample size, the method of Hudson (1985) may be competitive with the Monte Carlo Markov chain methods. (For small sample sizes, the point may be moot, since the sample probabilities have already been calculated and tabulated, as described in results and applications. For larger sample sizes, the issue remains important.) We now describe the method of Hudson (1985).

A two-locus genealogy is produced by generating a random sequence of events, proceeding backward in time from the present, as described by Hudson (1983). The events are coalescent events, in which two lineages merge into a single common ancestor, and recombination events, in which a single ancestral chromosome splits into two parental chromosomes. We denote the ith event by E_i. The complete ordered sequence of events is denoted $\boldsymbol{\epsilon}$ and is referred to as the E-sequence. Associated with each event is a specification of which lineage or lineages are involved. The E-sequence completely determines the topology of the A locus and B locus gene trees. A complete specification of the two-locus genealogy requires that one also specify the time intervals between the events. We note, however, that under the constant population size model, the E-sequence can be generated without regard to the time intervals between events. The time interval preceding E_i is denoted T_i. The ordered sequence of these time intervals is referred to as the T-sequence. Under the constant population size model and conditional on the E-sequence, the time intervals are independent exponentially distributed random variables. The mean of T_i depends on the configuration of the ancestral lineages during the interval, which, in turn, depends on the E-sequence. The calculation of the mean of T_i conditional on the E-sequence is also described in Hudson (1983).

The two-locus genealogy can be summarized as two tip-labeled gene trees, one for the A locus and one for the B locus. We arbitrarily number the branches of the A locus tree from 1 to $2n - 2$ and designate the length of the ith branch by a_i, measuring time in units of $4N$ generations. Note that for any particular branch, say the ith, a_i is the sum of one or more consecutive elements of the T-sequence. Similarly, the branches of the B locus tree are numbered, and their lengths are denoted by b_j. As with the A locus tree, the lengths of the B locus tree are sums of one or more consecutive elements of the T-sequence. It is assumed that the number of mutations on branch i of the A locus tree, conditional on its length, is Poisson distributed with mean $(\theta/2)a_i$. The sum of the a_i is denoted τ_A and the sum of the lengths of the B locus branches by τ_B. For a given two-locus genealogy, it is a simple matter to check for each pair of branches, one from the A locus gene tree and one from the B locus gene tree, whether a mutation on the A locus branch and a mutation on the B locus branch would lead to the specified sample configuration, $\mathbf{n}$. This property of the two-locus genealogy depends only on $\boldsymbol{\epsilon}$ and not on the T-sequence. Let $I(\boldsymbol{\epsilon}, \mathbf{n}, j, k)$ denote an indicator variable that is one if branch j on the A locus tree and branch k on the B locus tree are such a pair of branches and zero otherwise. If $I(\boldsymbol{\epsilon}, \mathbf{n}, j, k)$ equals one, then the sample configuration, $\mathbf{n}$, would arise if one or more mutations occurred on branch j of the A locus tree, and one or more mutations occurred on branch k of the B locus tree, and no mutations occurred elsewhere on the tree. Thus given $\boldsymbol{\epsilon}$ and the T-sequence, the probability of the configuration $\mathbf{n}$ being produced by mutations on branch j of the A locus tree and branch k of the B locus tree is

$$I(\boldsymbol{\epsilon}, \mathbf{n}, j, k)(1 - e^{-\theta a_j})(1 - e^{-\theta b_k})e^{-\theta(\tau_A - a_j)}e^{-\theta(\tau_B - b_k)}. \qquad (1)$$

Thus to obtain the sample probability, $q_u(\mathbf{n}; \theta, \rho)$, we sum over all branches j and k and take the expectation over the joint distribution of $\boldsymbol{\epsilon}$ and the T-sequence,

$$\begin{aligned} q_u(\mathbf{n}; \theta, \rho) &= E\left(\sum_{j,k} I(\epsilon, \mathbf{n}, j, k)(1 - e^{-\theta a_j})(1 - e^{-\theta b_k})(e^{-\theta(\tau_A - a_j)}e^{-\theta(\tau_B - b_k)})\right) \\ &\approx E\left(\theta^2 \sum_{j,k} I(\epsilon, \mathbf{n}, j, k)a_j b_k\right), \qquad (2) \end{aligned}$$

where $E(\;)$ designates expectation over random genealogies, j indexes the branches of the A locus tree, and k indexes the branches of the B locus tree. The approximation is for small θ and is obtained by Taylor expanding the exponentials and dropping higher-order terms in θ. Since we are interested in small θ, we consider the following function:

$$\begin{aligned} h_u(\mathbf{n}, \rho) &= \lim_{\theta \to 0} q_u(\mathbf{n}; \theta, \rho)/\theta^2 \\ &= E\left(\sum_{j,k} I(\boldsymbol{\epsilon}, \mathbf{n}, j, k)a_j b_k\right). \qquad (3) \end{aligned}$$

This function is perhaps best described as the "scaled, small-θ, likelihood function" and is referred to as the "scaled likelihood." The value of $h_u(\mathbf{n}, \rho)$ at a particular value of ρ can be estimated by generating a large number, m, of two locus genealogies using the specified value of ρ and calculating the sum

$$\widehat{h_u}(\mathbf{n}, \rho) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k} I(\boldsymbol{\epsilon}_i, \mathbf{n}, j, k)a_j(i)b_k(i), \qquad (4)$$

where $\boldsymbol{\epsilon}_i$ is the E-sequence of the ith randomly generated two-locus genealogy, and $a_j(i)$ and $b_k(i)$ are the branch lengths on the same two-locus genealogy. In effect, the method simply estimates the expected product of the lengths of pairs of branches that, if mutations occurred on them, would produce the specified sample configuration. To obtain an estimate of $q_u(\mathbf{n}; \theta, \rho)$ we use $\theta^2\widehat{h_u}(\mathbf{n}, \rho)$. This is the method of Hudson (1985).

In the case of the constant population size model, the method of Hudson (1985) can be made more efficient by replacing the randomly generated values of $a_j b_k$ by their expectation conditional on the E-sequence. That is, we estimate $h_u(\mathbf{n}, \rho)$ by

$$\widehat{\widehat{h_u}}(\mathbf{n}, \rho) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k} I(\boldsymbol{\epsilon}_i, \mathbf{n}, j, k)E(a_j b_k | \boldsymbol{\epsilon}_i). \qquad (5)$$

This is feasible because a_j and b_k are sums of one or more consecutive elements of the T-sequence, which are exponentially distributed. If a_j and b_k share no elements of the T-sequence in common, then the expectation of the product is the product of the expectations. If they have elements in common, the expectation of the product is the product of the expectations plus the sum of the expectations of the elements that are common to both. For example, if a_j is equal to the sum of $T_2 + T_3$ and b_k is $T_3 + T_4$, then under the constant population size model, the expectation of a_jb_k is

$$E((T_2 + T_3)(T_3 + T_4)) = (\lambda_2 + \lambda_3)(\lambda_3 + \lambda_4) + \lambda_3,$$

where $\lambda_i = E(T_i|\epsilon)$. This follows from properties of the exponential distribution. Thus if $h_u(\mathbf{n}; \rho)$ is estimated with (5) rather than (4), the T-sequence does not need to be generated and a lower variance estimate of $h_u(\mathbf{n}; \rho)$ is obtained.

Ancestral and derived alleles: In the previous paragraphs, we did not specify which alleles were ancestral and which were the mutant (or derived). It is now common to obtain sequence from a closely related species and infer which alleles are ancestral. The probabilities of sample configurations with specified ancestral/derived states are no more difficult to calculate than the unspecified configurations. A sample in which the ancestral/derived status of each allele is specified is referred to as an "a-d-specified" sample, and otherwise the sample is "a-d unspecified." The algorithm we just described can be modified to estimate the probabilities of a-d-specified samples by simply changing the indicator function used. Golding's recursion can also be modified to calculate a-d-specified sample probabilities. We use the convention for a-d-specified samples that A_0 and B_0 denote the ancestral alleles and A_1 and B_1 denote the mutant alleles. For a-d-specified samples, the quantities corresponding to $q_u(\mathbf{n}; \theta, \rho)$ and $h_u(\mathbf{n}; \rho)$ are denoted $q(\mathbf{n}; \theta, \rho)$ and $h(\mathbf{n}; \rho)$.

The a-d-unspecified probabilities can be obtained from the a-d-specified probabilities by summing four (or fewer) distinct a-d-specified sample probabilities, which result in the same unspecified sample configuration. That is, we can obtain the a-d-unspecified probabilities with

$$\begin{aligned} q_u((i, j, k, l); \theta, \rho) = [&q((i, j, k, l); \theta, \rho) + q((k, l, i, j); \theta, \rho) \\ &+ q((j, i, l, k); \theta, \rho) \\ &+ q((l, k, j, i); \theta, \rho)]/\pi(i, j, k, l), \end{aligned} \quad (6)$$

where

$$\pi(i, j, k, l) = \begin{cases} 4 & \text{if } i = j = k = l \\ 2 & \text{if } i = j \text{ and } k = l \text{ and } j \neq k, \text{ or if} \\ & \quad i = k \text{ and } j = l \text{ and } i \neq j, \text{ or if} \\ & \quad i = l \text{ and } j = k \text{ and } i \neq j \\ 1 & \text{otherwise.} \end{cases}$$

In RESULTS AND APPLICATIONS, we compare scaled-likelihood curves for one a-d-unspecified sample and the corresponding a-d-specified configurations. In that section we also address the issue of whether knowledge of which alleles are ancestral can improve estimates of ρ.

The method of Hudson (1985) can be extended to any neutral model in which the two-locus genealogy can be efficiently generated. In particular, simple island models of geographic structure and models with changing population size can be easily accommodated. A program to estimate $h(\mathbf{n}; \rho)$ using (4) under these models is available at http://home.uchicago.edu/~rhudson1.

Sequenced samples with two polymorphic sites: In the previous sections, the samples considered are assayed only at two sites. The intervening and flanking nucleotide sites may or may not be polymorphic. This is exactly the situation considered by Nielsen (2000). In contrast, we consider now the case where a set of contiguous sites are sequenced in each individual and all sites are therefore examined, and all polymorphisms in the sequenced segment are detected. Thus, full haplotype information is obtained for all sites in the segment. This is the situation considered by Griffiths and Marjoram (1996) and Kuhner *et al.* (2000). Nielsen (2000), Griffiths and Marjoram (1996), and Kuhner *et al.* (2000) all analyze the very difficult problem of estimating the probability of samples with arbitrary numbers of linked sites. In contrast, we now limit ourselves to the special case where only two sites are found to be polymorphic in the sample (and the rest are monomorphic), because, in this case, the random-genealogies method of Hudson (1985) can be easily extended to calculate these sample probabilities for a sequenced segment. This can be done as follows.

If the segment sequenced is L nucleotides long, we use an L-locus model, where each locus corresponds to a nucleotide site. We number the nucleotide positions from 1 (the leftmost site) to L(rightmost site.) Instead of a two-locus gene genealogy we must consider an L-locus gene genealogy. Each site has associated with it a gene genealogy or gene tree. The sum of the lengths of the branches of the gene tree of the ith site is denoted τ_i. We denote $\Sigma_{i=1}^{L} \tau_i/L$ by τ_{seq}. We designate the positions of the two polymorphic sites by x and y. The branches of the tree of site x and of the tree site y are numbered arbitrarily from 1 to $2n - 2$, and the length of branch j of the tree of site x is labeled $\tau_{x,j}$, and similarly $\tau_{y,k}$ denotes the length of the kth branch of the tree of site y. The two-locus sample configuration at sites x and y is denoted by $\mathbf{n}$, as before. Hudson (1983) describes how to generate an L-locus gene genealogy. The E-series must now include information on where each crossover event occurs along the segment. We focus on the case of small mutation rates, and so the infinite-allele model is still appropriate for each site. Let u_t denote the total mutation rate for the set of L sites and u_t/L the mutation rate per site. We denote $4Nu_t$ by θ_t. In this notation, if θ_t/L is small, the expected number of polymorphic sites is $\sim\theta_t E(\tau_{\text{seq}})$, and the probability of no polymorphic sites in the sample is $\sim E(e^{-\theta_t\tau_{\text{seq}}})$. As before, we define $\rho = 4Nr$, but in this case r is the recombination rate per generation between the leftmost and rightmost sites of the sequenced segment. The probability of the fully sequenced a-d-specified sample with two polymorphic sites is denoted $q_{\text{seq}}(\mathbf{n}, x, y; \theta_t, \rho)$ and is given by

$$q_{\text{seq}}(\mathbf{n},\ x,\ y;\ \theta_{\text{t}},\ \rho) = E\Bigg(\sum_{j,k} I_{x,y}(\boldsymbol{\epsilon},\ \mathbf{n},\ j,\ k)(1 - e^{-\theta_{\text{t}}\tau_{x,j}/L}) \times (1 - e^{-\theta_{\text{t}}\tau_{y,k}/L}) \times (e^{-\theta_{\text{t}}(\tau_{\text{seq}}-\tau_{x,j}/L-\tau_{y,k}/L)})\Bigg) \approx E\Bigg((\theta_{\text{t}}/L)^2\sum_{j,k} I_{x,y}(\boldsymbol{\epsilon},\ \mathbf{n},\ j,\ k)\,x_j y_k e^{-\theta_{\text{t}}\tau_{\text{seq}}}\Bigg), \tag{7}$$

where the approximation is obtained by expanding selected exponential terms and dropping terms of order $(\theta_{\text{t}}/L)^3$ and higher and where $I_{x,y}$ is an indicator function, as before, but in this case it depends on the gene genealogies of site x and of site y. $I_{x,y}$ is one if the jth branch of the tree of site x and the kth branch of the tree of site y are such that mutations on them lead to the given a-d-specified sample configuration $\mathbf{n}$. This expression for the probability of a sequenced sample is essentially the same as the expression for the two-locus configuration (2), except for the last term, $e^{-\theta_{\text{t}}\tau_{\text{seq}}}$. The analogue of $h(\mathbf{n};\ \rho)$ for sequence data, which we denote $h_{\text{seq}}(\mathbf{n},\ x,\ y;\ \rho,\ \theta_{\text{t}})$, is

$$h_{\text{seq}}(\mathbf{n},\ x,\ y;\ \rho,\ \theta_{\text{t}}) = \lim_{L\to\infty} q_{\text{seq}}(\mathbf{n},\ x,\ y;\ \theta_{\text{t}},\ \rho)/(\theta_{\text{t}}/L)^2, \tag{8}$$

where θ_{t} is assumed constant (and does not depend on L). This can be estimated by

$$\widehat{h}_{\text{seq}}(\mathbf{n},\ x,\ y;\ \rho,\ \theta_{\text{t}}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k} I_{x,y}(\boldsymbol{\epsilon}_i,\ \mathbf{n},\ j,\ k)\,x_j(i)\,y_k(i)\,e^{-\theta_{\text{t}}\tau_{\text{seq}}(i)}, \tag{9}$$

where $\boldsymbol{\epsilon}_i$ is the E-sequence of the ith randomly generated L locus genealogy, and $x_j(i)$ and $y_k(i)$ are the branch lengths on the trees of site x and site y, respectively. And $\tau_{\text{seq}}(i)$ is τ_{seq} for this same L locus genealogy.

In RESULTS AND APPLICATIONS we compare $h(\mathbf{n};\ \rho)$ and $h_{\text{seq}}(\mathbf{n},\ x,\ y;\ \rho,\ \theta_{\text{t}})$ to see how much the knowledge that there are no other polymorphisms between or near the focal pair of sites affects an inference about ρ. The estimates of $h_{\text{seq}}(\mathbf{n},\ x,\ y;\ \rho,\ \theta_{\text{t}})$ obtained as described here may be useful for checking other algorithms for estimating sequenced sample configuration probabilities.

Diploid samples: Up to this point, we have considered samples consisting of haplotypes. It would be useful to have sample probabilities analogous to $q(\mathbf{n};\ \theta,\ \rho)$ for the case of diploid samples. We show here how the probabilities of diploid samples can be expressed in terms of the haploid sample probabilities.

For diploids, with two alleles segregating at each locus, there are 10 distinct diploid genotypes. Often the phase of double heterozygotes is not determined directly, in which case there are only 9 distinguishable diploid genotypes. However, we begin by considering the case where the haplotypes constituting double heterozygotes are determined experimentally. In this case, the data can be represented by a 10-vector, $\mathbf{n_d} = (n_0,\ n_1,\ \ldots,\ n_9)$. We use n_0 to denote the number of coupling-phase double heterozygotes (A_0B_0/A_1B_1) and n_1 to denote the number of repulsion-phase double heterozygotes (A_0B_1/A_1B_0). The numbers of each of the other diploid genotypes are designated by n_i, $i = 2, \ldots, 9$. From the vector, $\mathbf{n_d}$, we can count the number of each of the four possible chromosomes. That is, the vector $\mathbf{n_d}$ maps unambiguously to the underlying haploid data configuration, which we denote by $\mathbf{n}(\mathbf{n_d})$. Under random mating, the probability of $\mathbf{n_d}$ is $q(\mathbf{n}(\mathbf{n_d});\ \theta,\ \rho)$ times the probability that $2n$ haploids of configuration $\mathbf{n}(\mathbf{n_d})$ when randomly paired produce $\mathbf{n_d}$. In symbols,

$$q_{\text{dip}}(\mathbf{n_d};\ \theta,\ \rho) = b(\mathbf{n_d},\ \mathbf{n}(\mathbf{n_d}))\,q(\mathbf{n}(\mathbf{n_d});\ \theta,\ \rho), \tag{10}$$

where $b(\mathbf{n_d},\ \mathbf{n}(\mathbf{n_d}))$ is the probability under random pairing of getting $\mathbf{n_d}$ from a haploid configuration $\mathbf{n}(\mathbf{n_d})$. By counting up the possible pairings, one finds that

$$b(\mathbf{n_d},\ \mathbf{n}) = \frac{n!}{\prod_{i=0}^{9} n_i!}\,\frac{n_{00}!\,n_{01}!\,n_{10}!\,n_{11}!}{(2n)!}2^{n_{\text{het}}}, \tag{11}$$

where n_{het} is the number of diploid individuals that are heterozygous at one or two loci and where n_i is the ith element of $\mathbf{n_d}$ and n_{ij}, $i, j = 0, 1$ are the elements of $\mathbf{n}$.

Now consider the case where the phase of double heterozygotes is not determined by the experimenter. In this case, we cannot observe n_0 or n_1, but we do observe their sum. We denote the sum by n_{dh}. We denote the diploid data set in this case by $\mathbf{n_{d-9}} = (n_{dh},\ n_2,\ \ldots,\ n_9)$. Given n_{dh}, the actual number of coupling phase double heterozygotes, n_0, could be any value from zero to n_{dh}. Each of these possible values corresponds to a different $\mathbf{n_d}$ configuration. We denote these possible $\mathbf{n_d}$ configurations by $\mathbf{n_d}(i,\ \mathbf{n_{d-9}})$, where $i = 0, \ldots, n_{dh}$. That is, if $\mathbf{n_{d-9}} = (n_{dh},\ n_2,\ \ldots,\ n_9)$, then $\mathbf{n_d}(i,\ \mathbf{n_{d-9}}) = (i,\ n_{dh} - i,\ n_2,\ \ldots,\ n_9)$. Then the probability of a diploid configuration $\mathbf{n_{d-9}}$ is obtained by summing up the probability of each of these mutually exclusive possible $\mathbf{n_d}$ configurations, as:

$$q_{\text{dip}-9}(\mathbf{n_{d-9}};\ \theta,\ \rho) = \sum_{i=0}^{n_{dh},} q_{\text{dip}}(\mathbf{n_d}(i,\ \mathbf{n_{d-9}});\ \theta,\ \rho). \tag{12}$$

Thus, with the haploid sample probabilities in hand, it is a simple matter to calculate diploid sample probabilities, using (10), (11), and (12).

Conditional probabilities: Most applications of these two-locus sampling distributions will focus only on pairs of sites in which both sites are polymorphic in the sample. This means that rather than $q(\mathbf{n};\ \theta,\ \rho)$, it will be useful to consider the probability of specific sample configurations conditional on two alleles segregating in the sample at each locus. That is, it is useful to consider the conditional probability

$$q(\mathbf{n},\ \theta,\ \rho|2\ \text{alleles at each locus}) = \frac{q(\mathbf{n};\ \theta,\ \rho)}{\sum_{\text{m}} q(\mathbf{m};\ \theta,\ \rho)}, \tag{13}$$

where the summation is over all configurations, $\mathbf{m}$, with two alleles at each locus. In the limit as θ tends to zero this becomes

$$\lim_{\theta\to 0} q(\mathbf{n},\, \theta,\, \rho | 2 \text{ alleles at each locus}) = \frac{h(\mathbf{n};\, \rho)}{\sum_{\mathrm{m}} h(\mathbf{m};\, \rho)}, \qquad (14)$$

which we can estimate without specifying θ. This conditional probability for small θ is denoted $q_c(\mathbf{n};\, \rho)$.

It may also be of interest to condition on other events. For example, one may wish to limit consideration to polymorphisms in which the rarer allele has frequency >0.05, or one may wish to condition on precisely the marginal allele frequencies observed. These are easily calculated by changing the summation in the denominator of the right-hand side of (14). Various conditional probabilities are utilized in the following sections.

RESULTS AND APPLICATIONS

Example sampling distributions: I have used the Hudson (1985) Monte Carlo algorithm (and Equation 4 or 5) to estimate $h(\mathbf{n};\, \rho)$ for all possible two-locus sample configurations (with exactly two alleles at each locus) for samples sizes of 20, 30, 40, 50, and 100 and a range of ρ values between 0 and 100. These are available at http://home.uchicago.edu/~rhudson1. The program used to estimate these quantities is also available at this site. The program actually generates multilocus gene trees and simultaneously estimates the sample probabilities for a range of recombination rates and all possible sample configurations simultaneously. The results for sample size 40 have been compared for several configurations to the results of solving the recursions of Golding (1984). No significant discrepancies were found. Thus two very different approaches using entirely independent computer code produced essentially the same values for the probabilities of a large number of sample configurations over a range of ρ values. In addition, the results for large recombination rates converge on the free recombination configuration probabilities that are easy to calculate with Ewens sampling distribution (Ewens 1972) and the assumption of independence of the two sites. These two results give considerable reassurance that the Monte Carlo program functions correctly. The program can also be used to estimate two-locus sample probabilities under the island model of spatial structure and under a model with recent exponential growth in population size.

Figure 1 shows some conditional sampling distributions for a sample of size 90. This figure shows the asymmetric U-shaped distribution of linkage disequilibrium, which is typical for low recombination rates, the broad distribution of linkage disequilibrium for values of ρ in the range of 5–20, and the unimodal and nearly normal distribution of linkage disequilibrium for large ρ. An application of these distributions is described in the next section.

The conditional probabilities $q_c(\mathbf{n};\, \rho)$ when plotted as functions of ρ are conditional-likelihood curves. Estimates of three such curves are shown in Figure 2. Application of these likelihood curves for estimating ρ are described in a subsequent section, *Estimating* ρ.

Before proceeding to some applications of these sample probabilities we consider briefly the effects of knowing which alleles are ancestral and the effects of having full sequence data *vs.* assessing the variation at only two specified sites. In Figure 3, we show plots of $h(\mathbf{n};\, \rho)$ for a set of four a-d-specified samples and $h_u(\mathbf{n};\, \rho)$ for the corresponding a-d-unspecified sample. In the plot, we see that different a-d-specified samples, each corresponding to the same a-d-unspecified sample, can have very different likelihood curves. In this case, two of the configurations have monotonically decreasing likelihood curves, and the other two configurations have a maximum for intermediate values of ρ. However, two of the configurations, which have similar curves, have much higher probabilities than the other two configurations. Thus, although knowledge of which alleles are ancestral may in specific instances have an important impact on inferences about ρ, on average, knowledge of which alleles are ancestral may not provide very much more information. This suggestion is supported by the results concerning asymptotic variances of estimates of ρ using many pairs of sites, which is described later.

The effects of having full sequence data *vs.* assessing the variation at only two polymorphic sites are illustrated in Figure 4, in which we show plots of $h(\mathbf{n};\, \rho)$ and $h_{\mathrm{seq}}(\mathbf{n},\, x,\, y;\, \rho,\, \theta_t)$ as functions of ρ for $\mathbf{n} = (15, 15, 9, 1)$ and $\theta_t = 0.6$. In this case, the curves are very similar in shape. One should be cautious about generalizing from this particular result, but it appears that, for the case where only two sites are polymorphic, likelihoods based on full sequence data may be quite similar to the likelihood based on assays of only the pair of sites that are polymorphic. For other values of θ_t this may not hold.

Assessing observed levels of linkage disequilibrium: With the conditional probabilities described above, it is possible to assess whether or not the level of linkage disequilibrium observed between a particular pair of sites is compatible with our neutral model and a specified value of ρ. For example, suppose that we have a sample of 90 gametes in which two sites, 1000 bp apart, are assayed and found to be polymorphic, with sample configuration, $\mathbf{n} = (53, 7, 17, 13)$. (This is the sample configuration corresponding to $n_{11} = 13$ in Figure 1.) Note that the marginal allele frequencies of the derived alleles are $30(= 17 + 13)$ at one locus and $20(= 7 + 13)$ at the other locus. We ask the question of whether this observed configuration is compatible with the hypothesis that $4Nr_{\mathrm{bp}}$ $(= \rho_{bp})$ equals, say, 0.001, where r_{bp} is the recombination rate per base pair per generation. With these assumptions, the relevant recombination parameter for our pair of sites is $\rho = \rho_{\mathrm{bp}}\ 1000 = 1.0$. To assess whether the linkage disequilibrium observed is

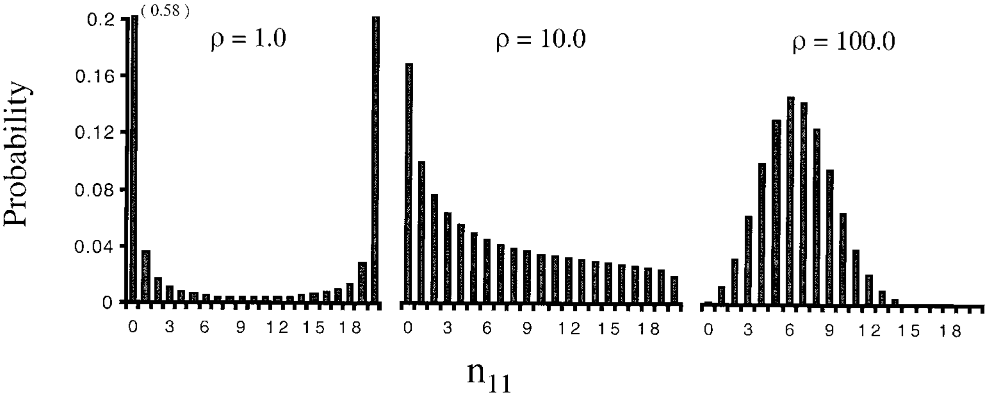

FIGURE 1.—Conditonal probabilities of two-locus sample configurations for samples of size 90. The height of each column gives the probability of the configuration with n_{11} taking the value on the abscissa, conditional on $n_{11} + n_{10} = 30$ and $n_{11} + n_{01} = 20$. Given these marginal frequencies, D takes its minimum possible value when $n_{11} = 0$ and its maximum possible value when $n_{11} = 20$. These conditional sample probabilities were calculated with (14) with the denominator equal to the sum over all configurations with the specified marginal allele frequencies. The leftmost column is truncated and should rise to a value of 0.58.

unusually high or low, we examine the distribution of **n** conditional on the observed marginal allele frequencies with $\rho = 1.0$. Conditional on the marginal allele frequencies, there are only 21 possible sample configurations, which can be specified by the value of n_{11}. The conditional probabilities of these configurations for $\rho = 1.0$ are shown in Figure 1 and were obtained using Equation 14, with the summation in the denominator of the right-hand side being over the 21 possible sample configurations.

We define a statistical test by summing the conditional probabilities of all configurations with probabilities less than or equal to the probability of the observed configuration. If this sum is <0.05 we reject our hypothesis. In our example, the configurations with $n_{11} = 8, \ldots,$ 13 have probabilities less than or equal to the observed configuration. The sum of the probabilities of these configurations is approximately 0.028, so our hypothesis is rejected. That is, we conclude that this sample configuration is quite unusual for $\rho = 1.0$ and note that it would be much more likely if $\rho = 10$. We refer to this test as the "exact test conditional on the marginals" (ETCM) and emphasize that it requires that one know or specify ρ.

Suppose now that our pair of polymorphic sites, with **n** = (53, 7, 17, 13), had been 100,000 bp apart, in which case, $\rho = 100$. If we examine the conditional probabilities for this value of ρ (shown on the right in Figure 1), we find that the sum of the probabilities of configurations with less or equal probabilities is ~0.02, and so the hypothesis would again be rejected. (In this case the configurations with lower probabilities are $n_{11} = 0$, and $n_{11} = 14, \ldots, 20$.) There is too much linkage disequilibrium in this case. This illustrates that the

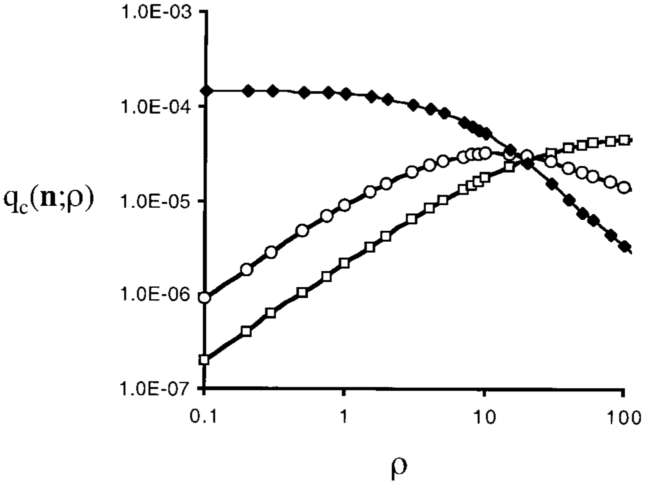

FIGURE 2.—The conditional-likelihood curves for three sample configurations. (□) $n_{11} = 5$. (○) $n_{11} = 1$. (◆) $n_{11} = 0$. The conditioning in this case is that there are two alleles at each locus in the sample. The three sample configurations are **n** = (20, 10, 20, 0), **n** = (21, 9, 19, 1), and **n** = (25, 5, 15, 5), which are labeled $n_{11} = 0$, $n_{11} = 1$, and $n_{11} = 5$, respectively. The marginal allele frequencies are the same for each configuration. The values of $q_c(\mathbf{n};\rho)$ shown here were estimated with (14), modified for a-d-specified samples, using scaled likelihoods estimated from (4).

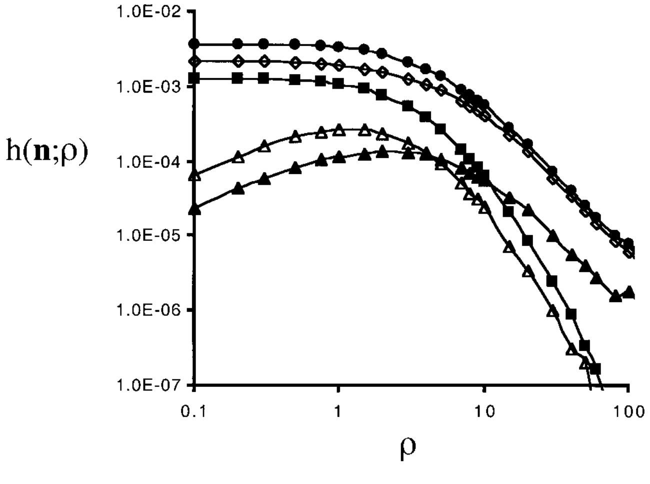

FIGURE 3.—A comparison of the scaled-likelihood functions for an a-d-unspecified sample, $h_u(\mathbf{n};\rho)$, and the four corresponding a-d-specified samples. The top curve, $h_u(\mathbf{n};\rho)$, is equal to the sum of the lower four curves. (●) h_u ((16, 20, 14, 0); ρ). (◇) h ((16, 20, 14, 0); ρ). (■) h((6, 30, 14, 0); ρ). (△) h ((0, 30, 14, 6); ρ). (▲) h((0, 20, 14, 16); ρ).

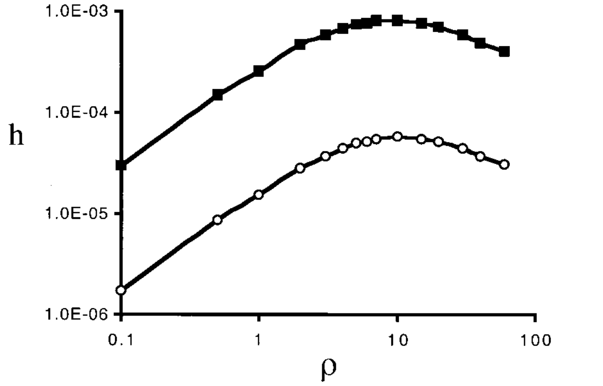

FIGURE 4.—A comparison of (■) $h(\mathbf{n}; \rho)$ and (○) $h_{\text{seq}}(\mathbf{n}, x, y; 2\rho, \theta_t)$ for $\mathbf{n} = (15, 15, 9, 1)$ and $\theta_t = 0.6$. $h_{\text{seq}}(\mathbf{n}; x, y; 2\rho, \theta_t)$ was estimated with Equation 9 on the basis of simulations with $L = 10{,}000$ sites and $x = 2500$ and $y = 7500$. With this choice of x and y, the recombination parameter corresponding to the recombination rate between these two sites is ρ.

ETCM can reject the null hypothesis due to either too much or too little linkage disequilibrium.

This test could be carried out for all possible pairs of polymorphic sites in a contiguous region to explore the possibility that some sites exhibit unusually high or low levels of linkage disequilibrium. Such sites may be indicative of hotspots of recombination or mutation or epistatic selection. This is a complementary approach to the usual analysis of linkage disequilibrium in which Fisher's exact test of independence is applied to all pairs. It should be noted that, when more than two linked sites are considered, there is a statistical dependence between the ETCMs on each pair, and any interpretation of the results should bear this in mind.

Estimating ρ: *Using independent linked pairs:* The conditional probabilities $q_c(\mathbf{n}; \rho)$ when plotted as functions of ρ are likelihood curves. Estimates of three such curves are shown in Figure 2. (The estimates are obtained using (14) and (4) but for a-d-specified samples.) Most sample configurations lead to monotonically increasing or decreasing likelihood curves, but samples with high but not complete linkage disequilibrium lead to likelihood functions with a maximum at a finite positive value of ρ. Thus the maximum-likelihood estimate of ρ for a single pair of sites is often zero or infinity. When the estimate is finite and greater than zero, the confidence interval is clearly large (as indicated by the broad likelihood function). This was noted before by HUDSON (1985) and by HILL and WEIR (1994). However, if one had k pairs of sites, where each pair is independent of the other pairs and where each pair of sites has the same ρ, then one might be able to obtain a very accurate estimate of ρ. In this case the overall likelihood, for small θ, is approximately

$$L(\mathbf{n}_1, \mathrm{n}_2, \ldots, \mathbf{n}_k; \rho) \approx \prod_{i}^{k} q_c(\mathbf{n}_i; \rho), \qquad (15)$$

in which $\mathbf{n}_i$ is the two-locus configuration for the ith pair of sites. The maximum-likelihood estimate of ρ obtained with (15) is denoted $\hat{\rho}$.

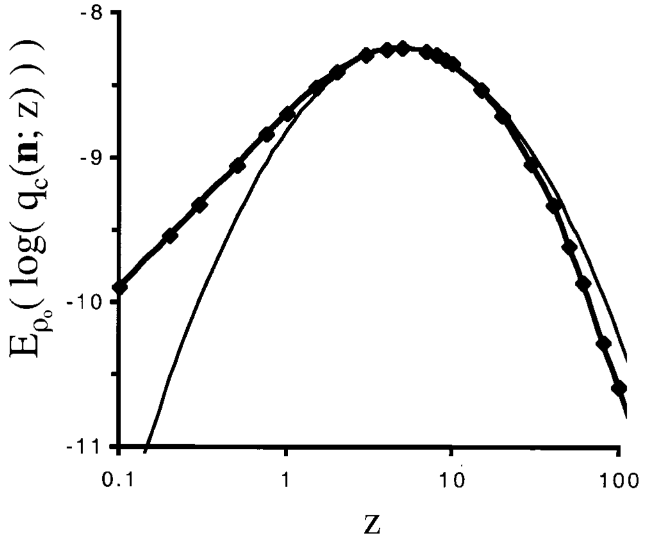

FIGURE 5.—The expected log-likelihood curve, (■) $E_{\rho 0}(\log(q_c(\mathbf{n}; z)))$, for $\rho_0 = 5.0$ and $n = 50$. For this curve we conditioned on the marginal allele frequencies being ≥0.1. (This curve is obtained using (16) and (14) and tabulated values of $h(\mathbf{n}; \rho)$.) Also shown is a second degree polynomial obtained by a least-squares fit to the points on the expected log-likelihood curve near $z = 5.0$. (—) Fitted quadratic.

To characterize the statistical properties of the maximum-likelihood estimate $\hat{\rho}$, it is useful to consider the expectation of $\log(q_c(\mathbf{n}; z))$, over the distribution of **n** conditional on polymorphism at both sites. That is, we consider

$$E_{\rho 0}(\log(q_c(\mathbf{n}; z))) = \sum_{\mathbf{n}} q_c(\mathbf{n}; \rho_0)\log(q_c(\mathbf{n}; z)), \qquad (16)$$

where $E_{\rho 0}$ indicates expectation given that the true value of ρ is ρ_0. This function can be viewed as a function of both z and ρ_0 and can be estimated from our tabulated values of $h(\mathbf{n}; \rho)$. An estimate of this function is plotted as a function of log z for $\rho_0 = 5.0$ and a sample of size 50 in Figure 5. (The conditioning for Figure 5 is that both sites are polymorphic with the rarer allele having frequency ≥0.1.) The second derivative of this function with respect to z, evaluated at the ρ_0, is inversely proportional to the asymptotic variance of the maximum-likelihood estimate of ρ. More precisely, if k pairs of sites are utilized, we expect the variance of the maximum-likelihood estimator to be approximately

$$\mathrm{Var}_{\rho_0,k}(\hat{\rho}) \approx \frac{1}{-k(\partial^2/\partial z^2)E_{\rho_0}(\log\, q_c(\mathbf{n}; z))|_{z=\rho_0}} \qquad (17)$$

for k sufficiently large. Here, $\mathrm{Var}_{\rho 0,k}$ denotes the variance of the estimator based on k pairs and with $\rho = \rho_0$. With the tabulated estimates of $h(\mathbf{n}; \rho)$, one can estimate the second derivative in (17) and hence the asymptotic variance of $\hat{\rho}$. However, it may be of most interest to

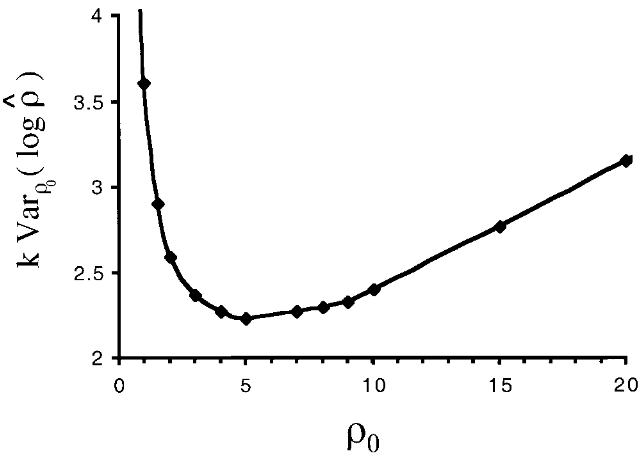

Figure 6.—Estimates of the asymptotic variance of the logarithm of the maximum-likelihood estimate of ρ based on *k* independent pairs of polymorphic sites. These estimates were obtained with Equation 18, with estimates of the second derivative of the expected log-likelihood function. The expected log-likelihood functions were estimated from (16) and tabulated values of $h(\mathbf{n}; \rho)$.

investigate the coefficient of variation of the estimate of ρ, so we consider instead

$$\frac{\mathrm{Var}_{\rho_0,k}(\hat\rho)}{\rho_0^2} \approx \mathrm{Var}_{\rho_0,k}(\widehat{\log(\rho)})$$

$$\approx \frac{1}{-k(\partial^2/\partial(\log z)^2)E_{\rho_0}(\log q_c(\mathbf{n}; z))|_{z=\rho_0}}. \qquad (18)$$

In Figure 5, we show, in addition to our estimate of $E_{\rho_0}(\log(q_c(\mathbf{n}; z)))$ for $\rho_0 = 5.0$, a quadratic function obtained by a least-squares fit to several points near $z = 5.0$. Clearly the expected likelihood function is very close to quadratic for a substantial range of *z* in the neighborhood of 5.0. This suggests that the asymptotic properties may apply for moderate values of *k*. By estimating the second derivative of the expected likelihood functions, we have estimated asymptotic variances (times *k*) for a set of values of ρ_0 and plotted the results as a function of ρ_0 in Figure 6. The plot in Figure 6 shows that pairs separated by ρ in the range of 2–15 are best for estimating ρ. As ρ decreases below 2.0, the asymptotic variance grows rapidly. For larger values of ρ the asymptotic variance grows more slowly.

To give some feeling for the number of pairs needed to get a reasonable estimate of ρ we consider a numerical example. Suppose that we have data for $k = 20$ independent pairs of polymorphic sites, where the rare allele has in every case allele frequency of at least 0.1 and where the recombination rate, ρ_0, between the sites of each pair is the same and is in the range from 2 to 10. In this case, we see from Figure 6 that $k\ \mathrm{Var}_{\rho_0,k}(\widehat{\log(\rho)})$ is ~2.5, and hence the asymptotic standard deviation of $\log(\hat\rho)$, when estimated from 20 independent pairs, is ~0.35 $(= \sqrt{2.5/20})$. If $\log(\hat\rho)$ is approximately normally distributed, then with probability ~0.95, $\log(\hat\rho)$ will lie in the interval $\log(\rho_0) \pm 2(0.35)$, and hence $\hat\rho$ will be within a factor of 2 of the true value. To check this result, I generated 32,000 two-locus samples on the computer, using the conditional sampling probabilities $q_c(\mathbf{n}; \rho)$, conditioning on the appropriate marginal allele frequencies. These random two-locus samples were formed into 1600 groups of 20, and $\hat\rho$ was calculated for each group. From these outcomes, the estimated variance of $\log(\hat\rho)$ was 0.124, which is very close to the prediction from the asymptotic analysis $(2.5/20 = 0.125)$. The probability of being within a factor of 2 is estimated to be 0.96, also in very good agreement with the asymptotic analysis prediction of 0.95.

In practice, different pairs of polymorphic sites will be different distances apart and will have different recombination rates. In the case where the physical distance between each pair of sites was known and the recombination rate per base pair was the same for each pair of polymorphic sites, then the likelihood for *k* polymorphic pairs is approximately

$$L(\mathbf{n}_1, \mathbf{n}_2, \ldots, \mathbf{n}_k; \rho_{\mathrm{bp}}) \approx \prod_{i}^{k} q_c(\mathbf{n}_i; \rho_{\mathrm{bp}} d_i), \qquad (19)$$

where ρ_{bp} is $4Nr_{\mathrm{bp}}$, r_{bp} is the recombination rate per base pair, and d_i is the distance in base pairs between the *i*th pair of sites. This likelihood could be used to estimate ρ_{bp}. The results in the previous paragraph suggest that the lowest variance estimator of ρ_{bp} will be obtained if sites are separated by a distance such that ρ_{bp} times the distance is ~5. If r_{bp} varied from one pair of sites to the next, but the value of r_{bp} were known for each pair of sites, say from comparisons of physical and genetic maps, a similar likelihood could be used to estimate *N*, the effective population size.

Using the above method, we can estimate the asymptotic variance of ρ in a-d-unspecified samples and in diploid samples in which the phase of double heterozygotes is not determined. For the case of a-d-unspecified samples of 50 gametes in which $\rho = 5.0$, the asymptotic variance is estimated to be $2.2/k$, which is essentially the same as what we found for a-d-specified samples. Thus, there appears to be little if any gain in knowing which alleles are ancestral when the data consist of independent linked pairs of sites. A similar asymptotic analysis could be used to determine the optimum sample size when one can trade off sample size for number of pairs. We do not pursue that analysis here.

Finally we note that, when investigating pairs of polymorphic sites, the incorporation of gene conversion is straightforward. It is necessary only to establish an effective recombination rate as a function of distance that incorporates gene conversion, such as Andolfatto and Nordborg (1998) or Langley *et al.* (2000). The scaled-likelihood functions do not need to be recalculated. Frisse *et al.* (2001) recently estimated gene conversion rates in humans using this method.

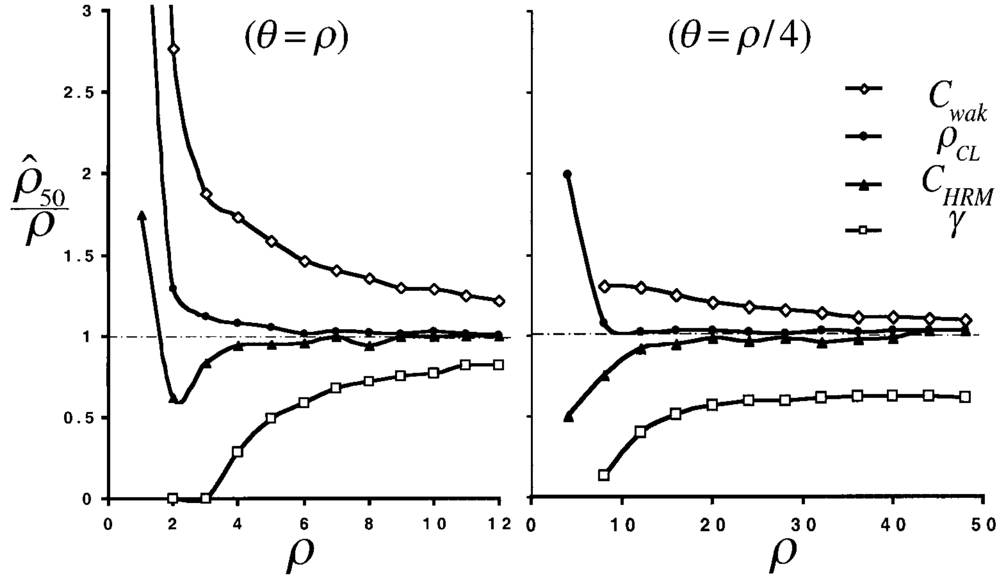

FIGURE 7.—Estimates of the medians ($\hat{\rho}_{50}$) of four estimators of ρ (each divided by the true value of ρ). These are based on 10,000 samples of size 50 generated by coalescent-based Monte Carlo simulations. The four estimators are described in the text.

Using many linked sites: In the previous sections, we considered one or more linked pairs of sites, where each pair is independent of the others. We now consider the case where more than two linked polymorphic sites are assayed in a single sample. In this situation, full likelihood considering all polymorphisms simultaneously is the proper approach. GRIFFITHS and MARJORAM (1996), KUHNER *et al.* (2000), and NIELSEN (2000) have provided Monte Carlo methods for estimating these likelihoods. However, at present the methods of GRIFFITHS and MARJORAM (1996) and KUHNER *et al.* (2000) are difficult to employ due to the very large computational requirements and the difficulty in determining when adequate convergence has been obtained. The approach of NIELSEN (2000) is less computationally demanding and goes some way toward solving this problem. However, the properties of full-likelihood estimators have not been explored, and the computation requirements for this exploration are daunting. As an alternative we consider a composite (or pseudo-) likelihood obtained by using (19), where the summation on the right-hand side is over all pairs of sites. This is an approach suggested by HUDSON (1993). Because of the statistical dependence between different pairs of sites, this expression is not the true likelihood. We can nevertheless maximize this function to obtain an estimate of ρ_{bp}, which is denoted $\hat{\rho}_{CL}$, where the subscript "CL" indicates an estimate based on composite likelihood. Once the two-locus sampling-scaled likelihoods ($h(\mathbf{n};\rho)$) are tabulated, calculating these composite likelihoods is very fast, and hence the statistical properties of this estimator can be explored.

To assess the quality of this composite-likelihood estimator, $\hat{\rho}_{CL}$ was calculated for a large number of samples generated by coalescent methods. The samples were generated by the method of HUDSON (1983), according to an infinite-site model, with ρ corresponding to the recombination rate between the ends of the segment observed and θ being the mutation parameter associated with the entire segment. Figures 7 and 8 show estimates of the medians and the 10th and 90th percentiles of the distribution of $\hat{\rho}_{CL}$ for a range of ρ values. The same quantities were estimated for three other estimators that have been described in the literature. These estimators are γ (HEY and WAKELEY 1997), C_{wak} (WAKELEY 1997), and C_{HRM} (WALL 2000). Each estimator was calculated for each of 10,000 samples (each of size 50 chromosomes). [These results for γ (HEY and WAKELEY 1997), C_{wak} (WAKELEY 1997), and C_{HRM} were kindly provided by Jeff Wall.]

The figures show that the estimator, C_{wak}, performs poorly compared to the other estimators shown. The estimator C_{wak} is an improved version of the estimator of HUDSON (1987), which would perform slightly more poorly than C_{wak}.

The estimator γ has a considerably lower 90th percentile than the other estimators. This is desirable as long as the 90th percentile is larger than the true value. Unfortunately, for large ρ and $\theta = \rho/4$, the 90th percentile of γ falls below the true value. In addition, it has a median considerably below the true value and a 10th percentile well below the 10th percentile of $\hat{\rho}_{CL}$ and C_{HRM}. Thus, for these parameter values γ has a strong tendency to underestimate ρ. For other parameter values HEY and WAKELEY (1997) showed that γ has little bias or a bias in the opposite direction.

The medians of the estimators $\hat{\rho}_{CL}$ and C_{HRM} are close to the true ρ, for $\rho > \sim 4.0$ for the case of $\theta = \rho$ and for $\rho > 10$ when $\theta = \rho/4$. The 10th percentiles of $\hat{\rho}_{CL}$ and C_{HRM} are considerably closer to the true value than the 10th percentiles of the other estimators. Their 90th percentiles are much closer to the true value than the 90th percentile of C_{wak} but as mentioned above are not as small as that of γ. In terms of percentiles, the estima-

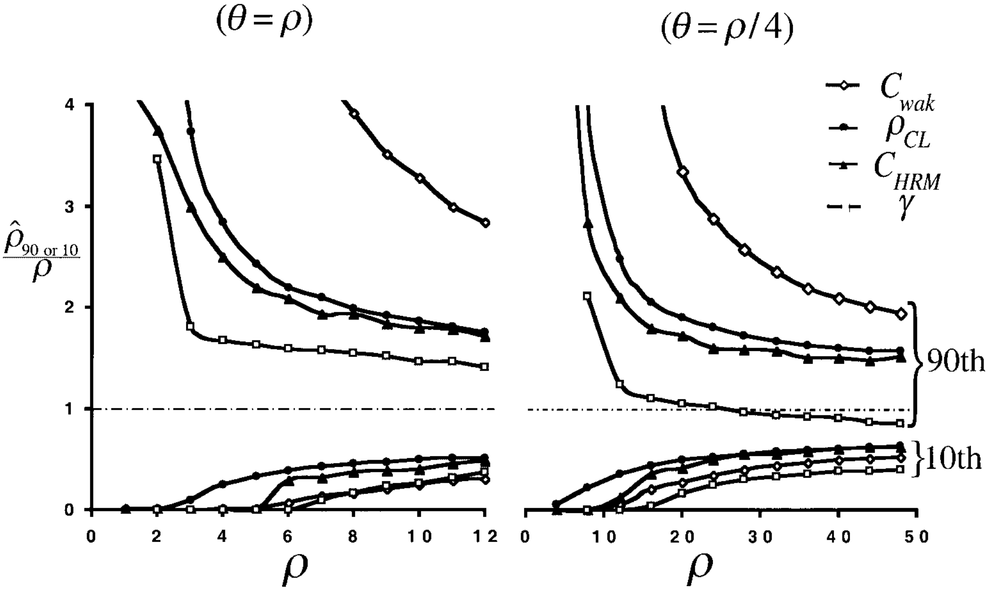

FIGURE 8.—Estimates of the 90th ($\hat{\rho}_{90}$) and 10th ($\hat{\rho}_{10}$) percentiles of the distributions of four estimators (divided by the true value of ρ). These were estimated with the same samples used in Figure 7.

tors $\hat{\rho}_{CL}$ and C_{HRM} appear to be quite similar. Overall, it appears that the estimators ρ_p and C_{HRM} are substantially superior to the other two estimators (at least for the parameter values investigated). The estimator $\hat{\rho}_{CL}$ has considerable flexibility that may make it of broader use. Since it does not rely on surveying or sequencing of all sites, it can be applied to data collected on previously identified single nucleotide polymorphisms (SNPs) or in surveys of regions with intervening gaps. Also, as indicated earlier, incorporating gene conversion is straightforward and does not require reestimating any of the $h(\mathbf{n}; \rho)$ values.

Finally, since $\hat{\rho}_{CL}$ is so quick to calculate (once the scaled likelihoods are in hand), one can afford to carry out simulations to characterize the properties of an estimate. For example, if the sampling procedure employed to collect the data is well specified and simple, then computer-generated samples can be used to study the distribution of the logarithm of the ratio of the composite likelihood of the data at $\hat{\rho}_{CL}$ to the likelihood at the true value of ρ for a range of ρ values and in this way obtain confidence intervals.

An application to human polymorphism data: We close by estimating ρ from a survey of human variation on the X chromosome (TAILLON-MILLER *et al.* 2000). In this study, 39 SNPs were surveyed in three population samples. In the following, only the sample of 92 CEPH males is considered. The parameter ρ_{bp} was estimated by maximizing the composite likelihood for (i) all 39 SNPs, (ii) the 14 SNPs in Xq25, and (iii) the 10 SNPs in or near Xq28. For loci on the X chromosome, using the $h(\mathbf{n}; \rho)$ functions described for autosomal loci will result in an estimate of $2Nr$, where r is the per-generation recombination rate in females and N is the total effective population size. In the following, the estimates returned by the computer programs were multiplied by 2 so that the values reported here for ρ are, in fact, estimates of $4Nr$, where r is the female recombination rate. The estimates were 9×10^{-5}, 8.8×10^{-5}, and 9×10^{-5} for all 39 sites, for the 14 sites of Xq25, and the Xq28 sites, respectively. These results suggest that there is no overall difference in recombination rates in the Xq25 region compared to the Xq28 region. The effective population size of humans has been estimated from levels of DNA polymorphism to be $\sim 10^4$ and the recombination rate per base pair, though quite variable, is for this region on average $\sim 10^{-8}$. Thus we might expect that $\rho_{bp} \approx 4 \times 10^{-4}$ or about five times larger than we estimate from the X chromosome data.

The linkage disequilibrium at each of the 741 ($= 39 \times 38/2$) possible pairs of sites was evaluated by the ETCM, described in *Assessing observed levels of linkage disequilibrium,* assuming $\rho_{bp} = 9 \times 10^{-5}$. A total of only 7 pairs, or $\sim$0.9% of the pairs, showed unusual two-locus configurations (with $P < 0.025$) using the ETCM. This is somewhat fewer than one would expect by chance when carrying out this many tests. Thus, overall, our analysis does not support the presence of a low recombination rate region in Xq25, as suggested by Taillon-Miller *et al.* However, it should be noted that 6 of the 7 significant pairs involve sites from the Xq25 region, and the seventh is immediately adjacent to the Xq25 region. Of the 6 significant pairs in the Xq25 region, 5 show unusually large linkage disequilibrium, but the 6th shows unusually low linkage disequilibrium. The latter pair of sites is separated by 30 kb and D' for the pair is 0.22. The five pairs showing significantly large linkage disequilibrium from the Xq25 region were among the pairs identified by Taillon-Miller *et al.* as showing significant linkage disequilibrium by a Fisher's exact test. In Figure 9, the values of r^2 are plotted for all pairs of sites within the Xq25 region (14 sites and hence 91 pairs) and for all

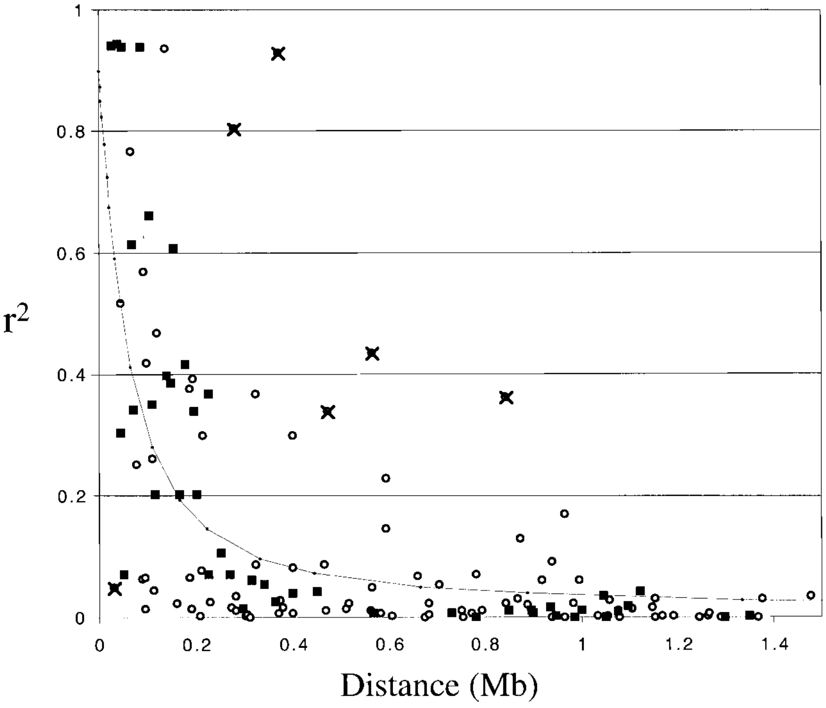

Figure 9.—A plot of r^2 *vs.* distance for polymorphic sites on the X chromosome in a CEPH sample of Europeans. The sites in the Xq25 region are indicated by open circles and those from the Xq28 region by solid squares. The points with x's over them (all from Xq25) have significantly unusual linkage disequilibrium by the ETCM test. The curve is the expected value of r^2 in samples of this size conditional on all alleles having frequency >0.32. See text.

pairs within the Xq28 region (10 sites or 45 pairs). Taillon-Miller *et al.* also displayed this plot. In the figure those sites that showed significantly large or small linkage disequilibrium given $\rho_{bp} = 9 \times 10^{-5}$, using the ETCM test, are shown with x's. Also shown in Figure 9 is the expected value of r^2 conditional on the frequencies of the minor alleles being >0.32. These were calculated with the tabulated $h(\mathbf{n}; \rho)$ values for a sample of size 92. The fit to the data appears to be fairly good, but there does appear to be some tendency for Xq28 pairs to fall below the line for larger distances and to be above the line for shorter distances. In contrast, the Xq25 sites appear to scatter on both sides of the curve.

CONCLUSIONS

Two-locus sample probabilities offer a useful tool for analyzing linkage disequilibrium levels from population surveys. Carrying out tests of significance for pairs of sites and estimating ρ from many sites is computationally quick, once the two-locus sample probabilities are in hand. A composite-likelihood approach for estimating ρ with more than two linked sites appears to work as well as the method of Wall (2000) and better than other *ad hoc* methods. Of course, none of these *ad hoc* methods should be used when full-likelihood methods are computationally feasible.

The author is grateful to Jeff Wall and Tony Long for useful discussion. Also, I am indebted to Jeff Wall for providing most of the numbers going into Figures 7 and 8 and to P. Taillon-Miller for providing the X chromosome data set. This research was supported in part by National Institutes of Health grants HG02098, HG01847, and HG02107.

LITERATURE CITED

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Chakravarti, A., K. H. Buetow, S. E. Antonarakis, P. G. Waber, C. D. Boehm *et al.*, 1984 Nonuniform recombination within the human beta-globin gene cluster. Am. J. Hum. Genet. **36:** 1239–1258.

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

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Ewens, W. J., 1979 *Mathematical Population Genetics.* Springer-Verlag, Berlin.

Frisse, L., R. R. Hudson, A. Bartoszewicz, J. D. Wall, J. Donfack *et al.*, 2001 Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. **69:** 831–843.

Golding, G. B., 1984 The sampling distribution of linkage disequilibrium. Genetics **108:** 257–274.

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

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of dna sequences with recombination. J. Comput. Biol. **3:** 479–502.

Hey, J., and J. Wakeley, 1997 A coalescent estimator of the population recombination rate. Genetics **145:** 833–846.

Hill, W. G., 1975 Linkage disequilibrium among multiple neutral alleles produced by mutation in a finite population. Theor. Popul. Biol. **8:** 117–126.

Hill, W. G., and B. S. Weir, 1994 Maximum-likelihood estimation of gene location by linkage disequilibrium. Am. J. Hum. Genet. **54:** 705–714.

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

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite allele model without selection. Genetics **109:** 611–631.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population mode. Genet. Res. **50:** 245–250.

Hudson, R. R., 1993 The how and why of generating gene genealogies, pp. 23–36 in *Mechanisms of Molecular Evolution,* edited by N. Takahata and A. G. Clark. Sinauer Associates, Sunderland, MA.

Karlin, S., and J. McGregor, 1968 Rates and probabilities of fixation for two locus random mating finite populations without selection. Genetics **58:** 141–159.

Kimura, M., and T. Ohta, 1971 *Theoretical Aspects of Population Genetics*. Princeton University Press, Princeton, NJ.

Kuhner, M. K., J. Yamato and J. Felsenstein, 2000 Maximum likelihood estimation of recombination rates from population data. Genetics **156:** 1393–1401.

Langley, C. H., 1977 Nonrandom associations between allozymes in natural populations of *Drosophila melanogaster*, pp. 265–273 in *Lecture Notes in Biomathematics, Vol. 19, Measuring Selection in Natural Populations*, edited by F. B. Christiansen and T. M. Fenchel. Springer-Verlag, New York.

Langley, C. H., Y. N. Tobari and K. Kojima, 1974 Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics **78:** 921–936.

Langley, C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen and J. M. Braverman, 2000 Linkage disequilibrium and the site frequency spectra in the *su(s)* and *su(w^a)* regions of the *Drosophila melanogaster X* chromosome. Genetics **156:** 1837–1852.

Lewontin, R. C., 1964 The interaction of selection and linkage. I. General considerations; heterotic models. Genetics **49:** 49–67.

Lewontin, R. C., 1974 *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York.

Lewontin, R. C., 1995 The detection of linkage disequilibrium in molecular sequence data. Genetics **140:** 377–388.

Macpherson, J. N., B. S. Weir and B. Leigh, 1990 Extensive linkage disequilibrium in the achaete-scute complex of *Drosophila melanogaster*. Genetics **126:** 121–129.

Nielsen, R., 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics **154:** 931–942.

Ohta, T., and M. Kimura, 1969 Linkage disequilibrium due to random genetic drift. Genet. Res. **13:** 47–55.

Ohta, T., and M. Kimura, 1971 Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68:** 571–580.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes in C*. Cambridge University Press, Cambridge, UK.

Taillon-Miller, P., I. Bauer-Sardina, N. L. Saccone, J. Putzel, T. Laitinen *et al.*, 2000 Juxtaposed regions of extensive and minimal linkage disequilibrium in human xq25 and xq28. Nat. Genet. **25:** 324–328.

Vieira, J., and B. Charlesworth, 2000 Evidence for selection at the fused locus of *Drosophila virilis*. Genetics **155:** 1701–1709.

Wakeley, J., 1997 Using the variance of pairwise differences to estimate the recombination rate. Genet. Res. **69:** 45–48.

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

Communicating editor: Y.-X. Fu