A Coalescent Estimator of the Population Recombination Rate

Jody Hey and John Wakeley

Department of Ecology, Evolution and Natural Resources, Nelson Laboratories, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received July 15, 1996
Accepted for publication November 7, 1996

ABSTRACT

Population genetic models often use a population recombination parameter $4Nc$, where N is the effective population size and c is the recombination rate per generation. In many ways $4Nc$ is comparable to $4Nu$, the population mutation rate. Both combine genome level and population level processes, and together they describe the rate of production of genetic variation in a population. However, $4Nc$ is more difficult to estimate. For a population sample of DNA sequences, historical recombination can only be detected if polymorphisms exist, and even then most recombination events are not detectable. This paper describes an estimator of $4Nc$, hereafter designated γ (gamma), that was developed using a coalescent model for a sample of four DNA sequences with recombination. The reliability of γ was assessed using multiple coalescent simulations. In general γ has low to moderate bias, and the reliability of γ is comparable, though less, than that for a widely used estimator of $4Nu$. If there exists an independent estimate of the recombination rate (per generation, per base pair), γ can be used to estimate the effective population size or the neutral mutation rate.

At the level of an individual DNA base pair, the source of all genetic variation is the process of mutation. However, at the level of a larger genotype such as the DNA sequence of a gene or of the genome of an organism, recombination is also a cause of genetic variation. Together these genome level processes, in conjunction with population level processes like genetic drift and natural selection, determine levels and patterns of genetic variation in natural populations.

In many population genetic models the conjunction of genome level and population level processes is complete, and the central parameter of the model is the product of a genome level (*e.g.*, mutation) rate and population level (*e.g.*, genetic drift) rate. One of the best examples is the neutral model prediction of the level of heterozygosity in a natural population. Under the simplifying assumptions of the neutral infinite-site model, the probability that a particular nucleotide site will be heterozygous in an individual from a diploid population is $4Nu$ (KIMURA 1969), where N is the effective population size, and u is the neutral mutation rate per base pair per generation. This particular parametric quantity (*i.e.*, $4Nu$) appears regularly in population genetic models and is usually referred to as θ (WATTERSON 1975; EWENS 1979). Similarly, the rate at which variation is generated by recombination is expected to depend on both N and c, where c is the rate of recombination per generation per base pair. For example, GRIFFITHS (1981) showed, under a two locus infinite-site model, that the covariance of the number of segregating sites at two loci is a function of both θ and $4Nc$. Hereafter, C will refer to $4Nc$.

While there exist several ways to estimate θ from comparative DNA sequence data (TAJIMA 1993; KUHNER *et al.* 1995), estimation of C is more difficult. In general, detection of recombination events in the history of a sample of DNA sequences depends on the amount of variation in the sample, and most recombination events are not detectable (HUDSON and KAPLAN 1985). HUDSON developed an estimator of C that is based on the variance of the number of base pair differences between DNA sequences (HUDSON 1987). However, HUDSON's estimator has low reliability if data sets are not very large (HUDSON 1987). In this report we describe a new estimator of C that works well over a wide range of sample sizes.

THEORY

Consider four DNA sequences drawn randomly from a population of constant effective population size. Assume that all mutations, and thus all polymorphisms among the four sequences, are neutral. In the time since the common ancestral sequences of all four sequences, there may have been many mutation and recombination events. We assume that the recombination and mutation rates are sufficiently low so that we can overlook the occurrence of multiple events (*i.e.*, more than one recombination event between adjacent base pairs or more than one mutation event at a base position) in the time since the common ancestors.

There are three possible unrooted topological structures for the genealogy of four DNA sequences at a particular DNA base position site (Figure 1). Each to-

Corresponding author: Jody Hey, Department of Ecology, Evolution and Natural Resources, Nelson Biological Labs, Rutgers University, Piscataway, NJ 08855-1059. E-mail: hey@mbcl.rutgers.edu.

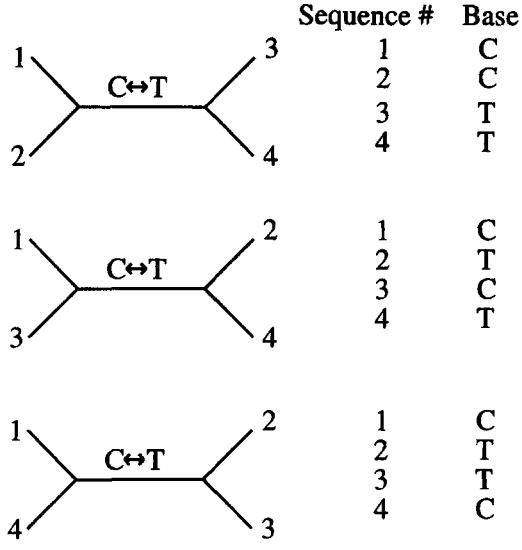

FIGURE 1.—Drawings of all three possible unrooted topologies for four items. The effect of internal branch mutations are shown as three different base patterns among the sequences.

pology can be distinguished by the particular pairs of sequences that are joined by the internal branch. A site becomes "informative" only if a mutation has occurred on the internal branch of that site's genealogy. A mutation on the internal branch will necessarily cause two of the sequences to show one nucleotide and the other two sequences to show a different nucleotide at the site of the mutation. Figure 1 shows how each topology, in conjunction with a mutation on the internal branch, corresponds to just one of three possible informative patterns.

When a pair of nucleotide sites is considered in a sample of four sequences, and there has not been recombination in the time since common ancestry, then both sites will have the same genealogical topology ("congruent" genealogies). If both sites are informative, then both will have the same pattern of base values (Figure 1). If recombination has occurred between the two sites in their history since the time of ancestral sequences, then it is possible for the two sites to have different genealogical topologies ("incongruent" genealogies). For the recombination to be detectable, it is necessary that both sites also be informative. If recombination has occurred and both sites are informative, then the two sites may reveal two of the patterns shown in Figure 1. This criterion for the detection of recombination has been called the "four-gamete" test (HUDSON and KAPLAN 1985).

Our goal is to develop a mathematical expression for the probability that a pair of informative sites, in a sample of four sequences, have congruent genealogies. The first step is to derive three quantities: I_2, I_1, and I_0. I_2 is the probability that two adjacent sites have incongruent genealogies given that they are both informative. I_1 is the probability that two adjacent sites have incongruent genealogies given that one of the sites is informative and that the other is not. I_0 is the probability that two adjacent sites have incongruent genealogies given that neither of them is informative. Although I_2 has been defined for adjacent sites, it is shown below how I_1 and I_0 can be used to generate an expression for $I_{2,n}$ which is the probability that two sites separated by n intervening sites have incongruent genealogies given that they are both informative (and that none of the intervening sites are informative).

Let the population recombination rate between two adjacent sites be $C = 4Nc$, where N is the effective population size and c is the recombination rate per gamete per generation between the two sites. Likewise, let the population mutation rate at each site be $\theta = 4Nu$, where u is the mutation rate per site per gamete per generation. The strategy in deriving I_2 is to calculate two probabilities:

$$P(\text{Genealogies Incongruent and Both Sites Informative}) \quad (1)$$

and

$$P(\text{Both Sites Informative}). \quad (2)$$

The quantity (2) is just the probability of there being two mutations, one on the internal branch of the genealogies of each of two adjacent sites. It can be calculated in three steps. First, enumerate all possible pairs of genealogies, and determine the probability of each. Second, calculate for each pair of genealogies the probability that both sites are informative, and third, determine the average probability of having both sites informative, calculated over all the possible pairs of genealogies and weighted by the probability of each pair of genealogies. Quantity (1) can be determined in exactly the same way by limiting the calculation to only those pairs of genealogies whose topological structures are incongruent.

Adopting a shorthand, (1) and (2) can be written as $P(\text{Incongruent and 2 Informative})$ and $P(\text{2 Informative})$, respectively. Then, using the standard expression for conditional probability,

$$I_2 = \frac{P(\text{Incongruent and 2 Informative})}{P(\text{2 Informative})}. \quad (3)$$

The expressions for I_1 and I_0 are similar, but rely on $P(\text{1 Informative})$ and $P(\text{0 Informative})$, respectively. In words, $P(\text{1 Informative})$ is the probability that one site is informative and the other site is not informative, and $P(\text{0 Informative})$ is the probability that neither of the two sites is informative. Using the formula for conditional probability,

$$I_1 = \frac{P(\text{Incongruent and 1 Informative})}{P(\text{1 Informative})} \quad (4)$$

and

$$I_0 = \frac{P(\text{Incongruent and 0 Informative})}{P(\text{0 Informative})}. \quad (5)$$

P(Incongruent and 1 Informative) and P(Incongruent and 0 Informative) are the same as P(1 Informative) and P(0 Informative), but summed over only those cases when the genealogies of the two sites are incongruent.

The genealogy of a particular site is defined by three common ancestor, or coalescent, events. If two adjacent sites remain linked during their entire history (*i.e.*, there is no recombination event between them), then the same three coalescent events define the trees at both sites and their genealogies are identical (and, of course, congruent). If there is a recombination event between the two sites, then their genealogies can differ and might be incongruent.

The pairs of genealogies for adjacent sites can be classified on the basis of the time of the recombination event. In a coalescent view, looking backward in time, the genealogies of two adjacent sites are identical up to the time when a recombination event occurred. If a recombination event does occur, it happens either (I) before the first coalescent event, (II) after the first but before the second coalescent event, or (III) after the second coalescent event and before the third one. If recombination does not occur before the first coalescent event, then the two adjacent sites will necessarily be congruent. That is, II and III above always result in congruent genealogies. In the case of class I, the two sites' genealogies can be either congruent or incongruent. No recombination event at all constitutes the fourth possibility, class IV. The probabilities of these four classes of genealogies are

$$P(\mathrm{I}) = \frac{C}{C+3}, \tag{6}$$

$$P(\mathrm{II}) = \left(\frac{3}{C+3}\right)\left(\frac{C}{C+2}\right), \tag{7}$$

$$P(\mathrm{III}) = \left(\frac{3}{C+3}\right)\left(\frac{2}{C+2}\right)\left(\frac{C}{C+1}\right), \tag{8}$$

$$\text{and} \quad P(\mathrm{IV}) = \left(\frac{3}{C+3}\right)\left(\frac{2}{C+2}\right)\left(\frac{1}{C+1}\right) \tag{9}$$

(HUDSON and KAPLAN 1985).

Within each class (I–IV), there are multiple genealogies to be considered. For example, Figure 2 shows the different genealogies that can occur for the case of recombination between the second and third coalescent events, class III, for two adjacent sites labeled A and B. Genealogies i and ii in Figure 2 have different topologies, but are identical in their contribution to I_2, I_1, and I_0. This is because the internal branch on which an informative change can occur is the same in both. Since, in all of i–iv the two site genealogies are congruent, P(Incongruent and 2 Informative), P(Incongruent and 1 Informative), and P(Incongruent and 0 Informative) are all equal to zero for each of the four

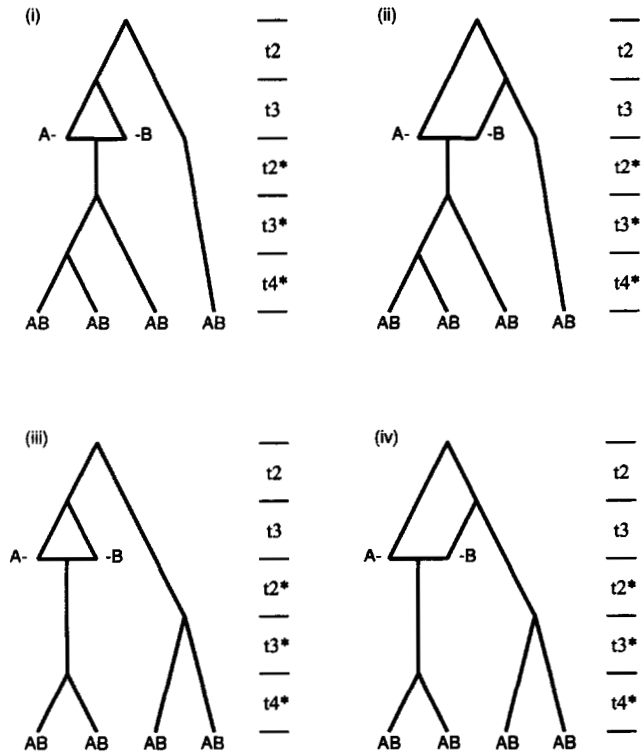

FIGURE 2.—Four distinct genealogies for the case when a recombination event occurs after two coalescent events, and before the third. A and B are adjacent nucleotide position. For each tree, one of two lineages that persists after two coalescent events becomes two lineages via recombination. Only those trees are shown in which the lineage on the left undergoes recombination at time t_2^*. Each of the trees i–iv has an equally likely counterpart in which the lineage on the right undergoes recombination. See text and APPENDIX for further explanation.

genealogies in the figure. The probabilities of genealogies i, ii, iii, and iv are 2/9, 4/9, 1/9, and 2/9, respectively (given class III), which can be determined by considering the numbers of different possible coalescent and recombination events that can happen in class III (TAJIMA 1983).

In calculating P(2 Informative), P(1 Informative), and P(0 Informative) for genealogies i–iv in Figure 2, the distributions of times between coalescent and recombination events shown must be specified. Looking back into the past, during some time intervals, both recombination and coalescent events might happen. In Figure 2, these are marked with asterisks. That is, t_i^* represents the time during which there are i ancestral lineages present and in which either a coalescent or a recombination event can occur. The density of t_i^*, $f(t_i^*)$ is given by an exponential distribution with parameter $ic + i(i-1)/4N$ (HUDSON 1983b).

Once a recombination event has occurred, no more are possible (under the assumptions of the model), so the history of the sample before the time of recombination is governed only by the coalescent process. Let t_i be the time during which there are i ancestral lineages when only a coalescent event may happen. Then the

density of t_i, $f(t_i)$ is given by an exponential distribution with parameter $i(i-1)/4N$ (KINGMAN 1982a,b; HUDSON 1983b; TAJIMA 1983).

Let the length of the internal branch of the genealogy of a site be τ. Given a particular value of τ, and assuming that u is small, the probability of a mutation that causes the site to be informative is simply $u\tau$. Similarly, the probability of no mutation is $(1-u\tau)$. If τ_A and τ_B are the lengths of the internal branches for the two sites, A and B, then

$$P(\text{2 Informative given } \tau_A \text{ and } \tau_B) = u^2\tau_A\tau_B. \quad (10)$$

This is because, for given values of τ_A and τ_B, mutation occurs independently at the two sites. The same property of independence leads to

$$P(\text{1 Informative given } \tau_A \text{ and } \tau_B) = u\tau_A(1-u\tau_B) + u\tau_B(1-u\tau_A) = u\tau_A + u\tau_B - 2u^2\tau_A\tau_B \quad (11)$$

and

$$P(\text{0 Informative given } \tau_A \text{ and } \tau_B) = (1-u\tau_A)(1-u\tau_B) = 1 - u\tau_A - u\tau_B + u^2\tau_A\tau_B. \quad (12)$$

In calculating the contributions to I_2, I_1, and I_0 for any particular tree, it is necessary to take the expectation of $u\tau_A$, $u\tau_B$, and $u^2\tau_A\tau_B$ over the distributions of the random variables τ_A and τ_B. The expectation of $u\tau_A$ is

$$u\int_0^\infty \tau_A f(\tau_A)\, d\tau_A, \quad (13)$$

where $f(\tau_A)$ is the probability density of τ_A. An analogous expression holds for $u\tau_B$. For $u^2\tau_A\tau_B$, the expectation over all possible values of τ_A and τ_B is

$$u^2\int_0^\infty\int_0^\infty \tau_A\tau_B f(\tau_A, \tau_B)\, d\tau_A\, d\tau_B, \quad (14)$$

where $f(\tau_A, \tau_B)$ is the joint distribution of τ_A and τ_B. These calculations are simplified by recognizing that τ_A and τ_B are sums of the times between branch points on the genealogies. Furthermore the t's and t^*'s are independent of one another. For example, from tree iv in Figure 2,

$$\tau_A = t_3^* + 2t_2^* + 2t_3 + 2t_2,$$

$$\tau_B = t_3^* + 2t_2^* + 2t_3,$$

and

$$\tau_A\tau_B = t_3^{*2} + 4t_2^*t_3^* + 4t_3^*t_3 + 4t_2^{*2} + 8t_2^*t_3 + 4t_3^2 + 2t_2t_3^* + 4t_2t_2^* + 4t_2t_3. \quad (15)$$

Then using the first and second moments of the exponential distributions for t_i and t_i^*, it can be shown that

$$E(u\tau_A) = \frac{\theta(2C+3)(2C+5)}{3(C+1)(C+2)},$$

$$E(u\tau_B) = \frac{\theta(C^2+7C+9)}{3(C+1)(C+2)},$$

and

$$E(u^2\tau_A\tau_B) = \frac{\theta^2(5C^4+50C^3+186C^2+299C+176)}{9(C+1)^2(C+2)^2}. \quad (16)$$

The APPENDIX shows $E(u\tau_A)$, $E(u\tau_B)$, and $E(u^2\tau_A\tau_B)$ for every possible history of two adjacent sites, for each of the four cases (I–IV) discussed above. Also shown in the APPENDIX are the probabilities, within each class, of each genealogy. The probabilities of the four classes are given by (6)–(9) above.

When the values given in the Appendix are averaged, weighted by their within-class probabilities and by the probabilities of the four classes, (10)–(12) give

$$P(\text{2 Informative}) = \theta^2 z, \quad (17)$$

$$P(\text{1 Informative}) = \theta - 2\theta^2 z, \quad (18)$$

$$P(\text{0 Informative}) = 1 - \theta + \theta^2 z, \quad (19)$$

where

$$z = \frac{177C^3+1091C^2+2234C+1800}{360(C+1)(C+2)(C+3)}. \quad (20)$$

For the calculation of the joint probabilities of incongruent and informative sites, only the last six genealogies under case I listed in the APPENDIX need be considered, as these are the only genealogies that lead to incongruent sites. Averaging over these six cases leads to the following:

$$P(\text{Incongruent and 2 Informative}) = \frac{\theta^2 C}{20(C+3)}, \quad (21)$$

$$P(\text{Incongruent and 1 Informative}) = \frac{\theta C(10-3\theta)}{30(C+3)}, \quad (22)$$

and

$$P(\text{Incongruent and 0 Informative}) = \frac{C(24-20\theta+3\theta^2)}{60(C+3)}. \quad (23)$$

Then, I_2, I_1, and I_0 are obtained, as in (3)–(5), using expressions (17)–(23):

$$I_2 = \frac{C}{z\,20(C+3)}, \quad (24)$$

$$I_1 = \frac{\theta C(10-3\theta)}{30(C+3)(\theta-2\theta^2 z)}, \quad (25)$$

and

$$I_0 = \frac{C(24 - 20\theta + 3\theta^2)}{60(C + 3)(1 - \theta + \theta^2 z)}. \quad (26)$$

Clearly I_2 does not depend on θ. This is not the case for I_1 and I_0. However, an examination of the functions for I_1 and I_0 revealed relatively simple and accurate approximations for small values of θ. Plotting expressions (25) and (26) over a wide range of values for C, and for θ in the range between 0 and 0.1, showed the functions to be nearly flat as they vary with respect to θ (results not shown). In short, under the assumption that the mutation rate is very low, both I_1 and I_0 are effectively invariant with respect to θ. Taking the limit as θ goes to zero, for expressions (25) and (26), leads to

$$I_1 \approx \frac{C}{3(C + 3)} \quad (27)$$

and

$$I_0 \approx \frac{2C}{5(C + 3)}, \quad (28)$$

respectively.

Together, expressions (24) – (26) cover all possible situations, with regard to whether sites have congruent genealogies, that can occur for a pair of adjacent sites. However, only expression (24), for I_2, can correspond to an observation. If two adjacent informative sites are observed in a sample of four DNA sequences, then it is possible to assess whether or not they are congruent (HUDSON and KAPLAN 1985). However if neither, or just one, site is informative, then we cannot know whether the true genealogies of a pair of sites are congruent.

With expressions for I_1 and I_0 it is possible to develop an expression for $I_{2,n}$, the probability that two sites separated by n bases are incongruent, given that both are informative and that there are no informative sites in the n intervening base positions. For convenience we assume that the probabilities associated with sites that are separated by multiple bases can be assessed by taking the products of probabilities for adjacent bases. This approach entails an assumption of independence among pairs of sites that does not, in fact, hold (see below).

For the case of two informative sites that are separated by one noninformative site, the genealogies of three base positions must be considered. Since each position has three possible unrooted topologies, there are a total of 27 possible unrooted three base topologies to be considered. However, topologies need only be distinguished with regard to whether adjacent base positions are congruent and with whether the first base position is congruent with the last base position. Let X, Y and Z refer to three different topologies as described in Figure 1. It does not matter which specific configuration each letter refers to, rather the notation is used to describe whether sites are congruent with one another. In general we will use X to refer to a configuration, Y to refer to just one of the other configurations that is not the same as X, and Z to refer to the third configuration that is neither X nor Y. In the case of two informative sites separated by one noninformative site, one possible configuration is that the first and second bases are congruent, and that the third base is not congruent with the second (and thus also not congruent with the first). This pattern is denoted XXY and the probability of this pattern (assuming independence) is simply the product of the probability of congruent genealogies between sites one and two (*i.e.*, $1 - I_1$) multiplied times the probability of incongruent genealogies X and Y between sites two and three (*i.e.*, $I_1/2$). This latter term is divided by two because the pattern XY specifies a specific incongruent pair (*i.e.*, XY and not XZ). More generally, the probability of either XXY or XXZ is $(1 - I_1)I_1$. There are just three different combinations of topologies that need be distinguished for the case of two informative sites separated by one noninformative site:

$$P(XXY \text{ or } XXZ) = (1 - I_1)I_1; \quad P(XYY \text{ or } XZZ) = I_1(1 - I_1); \quad \text{and} \quad P(XYZ \text{ or } XZY) = I_1^2/2. \quad (29)$$

The sum of these terms is the total probability that two informative sites, separated by one noninformative site, have incongruent genealogies:

$$I_{2,1} = \frac{I_1(4 - 3I_1)}{2}. \quad (30)$$

An expression for higher values of n can be developed by first focusing on just the noninformative bases that separate informative sites. Because the probability of incongruency is the same for any pair of adjacent sites, and because we will generally be applying these expressions to cases where n is large and where I_0 is small, we can borrow directly from a well known expression that is widely used in studies of evolutionary distance (JUKES and CANTOR 1969). There are three differences: the four character states of JUKES and CANTOR are replaced by three character states (*i.e.*, the configurations in Figure 1); time, under the JUKES and CANTOR model, is replaced by the number of intervals between noninformative bases; and the probability of mutation between individuals bases, under the JUKES and CANTOR model, is replaced by the probability of the XY pattern (*i.e.*, $I_0/2$) for two noninformative sites.

Let $X \cdots X_{n-1}$ refer to the case where both the left and right bases in a string of n noninformative bases are in the same configuration, regardless of which of the three configurations that is. The subscript $n - 1$ refers to the number of intervals between the bases. Then the analogue to the JUKES-CANTOR expression for the probability that the same character state is observed becomes, under the current model,

$$P_0(X \cdots X_{n-1}) = \tfrac{1}{3} + \tfrac{2}{3}e^{-3(n-1)I_0/2}. \quad (31)$$

TABLE 1

Flanking incongruent informative bases separated by noninformative bases

Congruent noninformative bases		Incongruent noninformative bases	
Pattern	Probability	Pattern	Probability
$XX..X_{n-1}Y$	$(I_1(1-I_1)/2)\cdot P_0(X..X_{n-1})$	$XX..Y_{n-1}Y$	$(1-I_1)(1-I_1)\cdot P_0(X..Y_{n-1})$
$XY..Y_{n-1}Y$	$(I_1(1-I_1)/2)\cdot P_0(Y..Y_{n-1})$	$XY..X_{n-1}Y$	$I_1I_1/4\cdot P_0(Y..X_{n-1})$
$XY..Y_{n-1}Z$	$(I_1/2)^2\cdot P_0(Y..Y_{n-1})$	$XX..Z_{n-1}Y$	$(1-I_1)/I_1/2\cdot P_0(X..Z_{n-1})$
		$XZ..X_{n-1}Y$	$I_1I_1/4\cdot P_0(Z..X_{n-1})$
		$XY..Z_{n-1}Y$	$I_1I_1/4\cdot P_0(Y..Z_{n-1})$
		$XZ..Y_{n-1}Y$	$(1-I_1)I_1/2\cdot P_0(X..Z_{n-1})$
$2\cdot\Sigma$	$\frac{I_1(4-3I_1)}{2}\cdot\left(\frac{1}{3}+\frac{2}{3}e^{-I_0(n-1)3/2}\right)$		$\left(1-\frac{I_1(4-3I_1)}{4}\right)\left(\frac{2}{3}-\frac{2}{3}e^{-I_0(n-1)3/2}\right)$

For each pattern, the left- and rightmost symbols refer to informative bases. In between the informative bases is a stretch of n noninformative bases. X, Y and Z each represent different informative site configurations as shown in Figure 1, however the specific configuration for each of X, Y and Z is arbitrary (see text). Each pattern describes a possible multibase configuration when two informative bases are separated by a stretch of noninformative bases. All possible configurations are represented with the exception that each pattern also includes an equally likely corresponding pattern with Y replaced by Z, and Z by Y. For this reason the terms in the last row are twice the sum of the probabilities in the respective columns. Expression (31) applies to the P_0 probabilities in the left column. One-half of expression (32) applies to the those in the right column (see text).

There are two patterns to be considered for the case when the left and rightmost bases are incongruent, $X\cdots Y_{n-1}$ and $X\cdots Z_{n-1}$. The probability that either occurs is simply one minus expression (31), or

$$P_0(X\cdots Y_{n-1} \text{ or } X\cdots Z_{n-1}) = {}^2/_3 - {}^2/_3 e^{-3(n-1)I_0/2}. \quad (32)$$

At times it is necessary to consider just one of the two incongruent patterns that can occur for a given configuration X (*e.g.*, $X\cdots Y_{n-1}$, but not $X\cdots Z_{n-1}$). In this case the probability is one half that in (32).

We can now consider additional flanking bases that are informative. Table 1 lists the possible configurations for n bases where the left and right bases are informative. Each configuration includes an internal configuration corresponding to the possibilities for a stretch of noninformative bases. The left column of terms in Table 1 corresponds to the case when no net configuration change occurs across the span of noninformative sites. The right column describes the possible configurations when a net change does occur across the span of noninformative bases. Collectively these expressions cover all possible configurations in which two informative sites separated by n intervening noninformative bases can occur. The overall probability that two informative bases separated by n noninformative bases are incongruent ($I_{2,n}$) is simply the sum of the two expressions in the last row of Table 1:

$$I_{2,n} = \frac{2}{3} - \frac{(3I_1-2)^2 e^{-(n-1)I_0 3/2}}{6}. \quad (33)$$

Expression (33) applies for all n greater than or equal to one. When n is equal to one, expression (33) simplifies to expression (30). This is because a single site is necessarily congruent with itself, so that when n is equal to 1, $P(X\cdots X_{n-1}) = 1$. Expression (33) is not defined for n equal to zero, however this is already given by I_2 in expression (24). Using the approximate expressions for I_1 and I_0 in (27) and (28), and including the special case of I_2 when n is equal to zero, (33) becomes

$$I_{2,n} \approx \frac{2}{3} - \frac{1}{6}\left(\frac{C}{C+3} - 2\right)^2 e^{[-3C(n-1)]/[5(C+3)]},$$

for $n \geq 1$; and

$$I_{2,n} = \frac{C}{z20(C+3)}, \quad \text{for } n = 0. \quad (34)$$

We have also developed an exact expression for $I_{2,n}$ that does not entail the assumptions associated with the differential equation approach used by JUKES and CANTOR (1969) (though the assumption of independence among sites still applies) and that does not use the approximations of expressions (27) and (28) (results available upon request). Numerous numerical comparisons between this exact expression and expression (34) show that (34) is a very good approximation (within 1% of the exact expression), except when distances are short (*e.g.*, $n < 5$) and θ is high (*e.g.*, >0.05 per base pair).

A plot of (34) for several values of C is shown in Figure 3. The limit of two-thirds is the expected consequence of three equally likely configurations that can occur for an informative site (Figure 1). Two informative sites with many recombination events between them will be congruent one-third of the time and incongruent two-thirds of the time.

Estimating *C*: The expression for $I_{2,n}$ can be used to estimate C. Consider, for example, a sample of four sequences with three informative polymorphic sites. Let

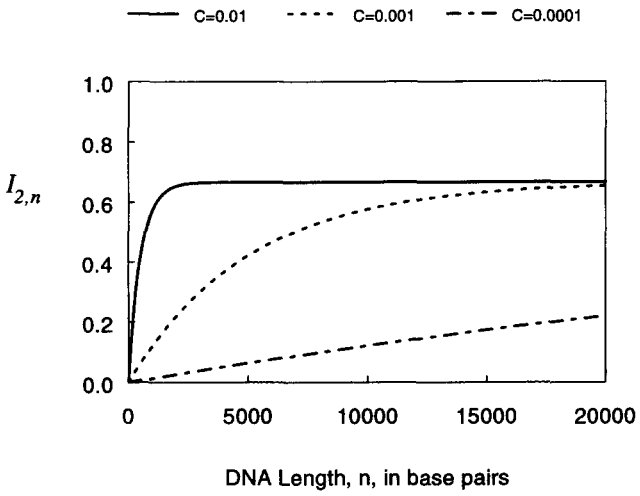

FIGURE 3.— $I_{2,n}$ calculated using expression (34) for three different levels of recombination, as n varies from 0 to 20,000.

the intervals between successive pairs of sites be numbered 1 and 2, with interval lengths n_1 and n_2. Each interval may be associated with sites that are either congruent or incongruent. Let $\psi_i = I_{2,ni}$ if interval i is incongruent, where $I_{2,ni}$ is calculated using expression (34), and let $\psi_i = 1 - I_{2,ni}$ if it is congruent. Then the overall likelihood of the pair of intervals can be expressed as

$$L = \psi_1\psi_2. \tag{35}$$

A maximum likelihood estimate of C is that value, which when substituted for C into (35), generates the largest value of L. For example, consider a hypothetical data set with three informative sites, in which one interval has a length of 100 bases and has congruent sites, while the second interval has a length of 500 bases and has incongruent sites. In this case, the likelihood function takes the form

$$L = (1 - I_{2500})\, I_{2100}. \tag{36}$$

By calculation or by plotting, this function can be shown to reach a maximal value at 0.452 when the estimated value of C is equal to 0.023.

When both intervals are congruent, there is no evidence of recombination. From a heuristic perspective, the most appropriate estimate of C would seem to be zero in this case. It can also be shown for the case of two congruent pairs of sites, at least with numerical examples, that the highest value of L, for nonnegative values of C, occurs when C is zero.

There are two situations in which (35) cannot be used to estimate C from a sample of four DNA sequences with three informative sites. If the data reveal only incongruent intervals, then they suggest a high recombination rate, but there is no way to put an upper bound on the estimate. In this case, numerical examples show that L approaches an asymptotic limit as C goes to infinity. It is also possible for this method to fail with one congruent and one incongruent interval. If by chance, the incongruent interval is short, relative to the length of the congruent interval, then L may not have a maximum for a finite value of C.

For data sets with more than four sequences and more than three informative sites, an overall estimate of C can be taken by averaging the estimate for every possible set of two intervals in all possible subsets of four sequences. For two intervals, i and j, in subsample k, let $\gamma_{(i,j)k}$ be an estimate of C obtained by maximizing the likelihood, as in (35). We denote the overall estimate by γ, which is equal to

$$\frac{\Sigma_k \dfrac{\Sigma_{i_k} \Sigma_{j_k \neq i_k} \gamma_{(i_k, j_k)}}{m_k}}{\binom{w}{4}}, \tag{37}$$

where i_k and j_k refer to two intervals within subsample k, and m_k is simply the total number of pairs of intervals found in subsample k. The final denominator, $\binom{w}{4}$, is just the number of distinct subsamples of size 4. When the number of sequences, w, is large, the number of possible subsamples becomes unwieldy. In this case γ can be calculated using multiple randomly drawn subsamples, each of size 4. When calculating γ, pairs of sites that do not generate a value of $\gamma_{(i,j)k}$ (*e.g.*, where both pairs are incongruent) are not included, and are not counted in m_k. Similarly, those subsets of four sequences for which no values of $\gamma_{(i,j)}$ could be calculated (*e.g.*, if the number of intervals in the subset is zero or one) are not counted, and the denominator of (37) is reduced by the number of such subsets.

There are two important assumptions employed in the development of γ that do not hold. One is inherent to the biology of DNA replication, and this is the assumption that the probability of incongruent histories for sites that are not adjacent is equal to the product of the probabilities for adjacent base pairs. This can only be true if adjacent base positions have independent histories (*i.e.*, free recombination), which is definitely assumed to not be the case under the model. In effect the method has conflicting assumptions: an assumption that adjacent base pairs do not have independent histories that is employed in developing expressions (24), (25) and (26); and an implicit assumption of independence that is employed to generate the expression for $I_{2,n}$. The same implicit assumption of independence appears in the use of the maximum likelihood method, where the likelihood for multiple sets of pairs of sites is determined by taking the product of the likelihood for each set.

A second implicit assumption that does not hold in practice is that the length of DNA sequences used in a study does not effect the probability of observing particular lengths for congruent or incongruent intervals. However neither $I_{2,n}$ nor $(1 - I_{2,n})$ are approximated very well for values of n that are close to the total length of the DNA sequence. If DNA sequences are short, then more intervals will be congruent simply because the

expected length of congruent intervals is less than that for incongruent intervals. The result will be a lower value of γ for data sets with short sequences, all other things being equal.

The quality of γ was examined using computer simulation. A computer program was written to implement a standard coalescent process with recombination (HUDSON 1983a). This program did permit multiple recombination events at a given position over the course of a sample coalescence. In this the program model corresponds more closely to biological reality than to the assumptions employed in estimating C. However, the difference should have negligible impact for low values of C.

Expression (37) requires either graphical or numerical maximization, and the summation across pairs of intervals and subsets of four DNA sequences can be tedious. A computer program, SITES, was written and used to analyze several comparative DNA sequence data sets drawn from the literature. SITES is a general purpose program for the analysis of comparative DNA sequence data, and it is intended primarily for cases when multiple sequences are collected from a population or species. SITES can be obtained via request to J.H.

RESULTS

The quality of γ: Simulations were conducted over a wide range of parameter values, for both population parameters (C and θ) and experimental parameters (*e.g.*, the number of DNA sequences and the length of DNA sequences). For each set of parameter values, many simulations were run. The bias of the estimator, under a particular set of parameter values, was assessed by taking the mean value of γ obtained from the simulations and dividing by the parametric value of C. This measure has a value of one when there is no bias. We also assessed the variation of γ around the true parametric value of C. We defined a quantity, ϵ, which is equal to the square root of the estimated mean square error (MSE), divided by the parametric value:

$$\epsilon = \frac{\sqrt{\widehat{\mathrm{MSE}}}}{C}, \quad \text{where} \tag{38}$$

$$\widehat{\mathrm{MSE}} = \frac{\sum_{i=1}^{R} (\gamma_i - C)^2}{R}, \tag{39}$$

and R is the number of simulations. In effect, ϵ is a measure of the spread in the distribution of departures of the estimates of C from the true value, as a proportion of the true value.

Figure 4a shows the bias and ϵ for γ, as a function of changing sample size. Also shown in Figure 4a are bias and ϵ for WATTERSON's estimator of θ (WATTERSON 1975; HUDSON 1990) and for HUDSON's (HUDSON 1987) estimator of C (Figure 4b) that were generated using the same simulations. The simulations for Figure 4 were done with parametric values of $C = 0.02$, and $\theta = 0.005$, for a DNA sequence length of 2000 base pairs. These values were selected because they are representative of results from analyses of two data sets from *Drosophila melanogaster* (see Applications). For these parametric values, γ has some bias (the mean value is about two-thirds of the parametric value) but the bias is not a function of the number of DNA sequences in the sample. For γ, ϵ is about three times larger than it is for $\hat{\theta}$ for a sample size of 4, and drops down to about twice that of $\hat{\theta}$ for larger sample sizes. Thus, despite a bias in the estimator, the variation in γ about the true value of C, as measured by ϵ, is only two to three times larger than it is for WATTERSON's estimator of θ. HUDSON's (1987) estimator of C is not reliable for small sample sizes (Figure 4b), but does approach the range

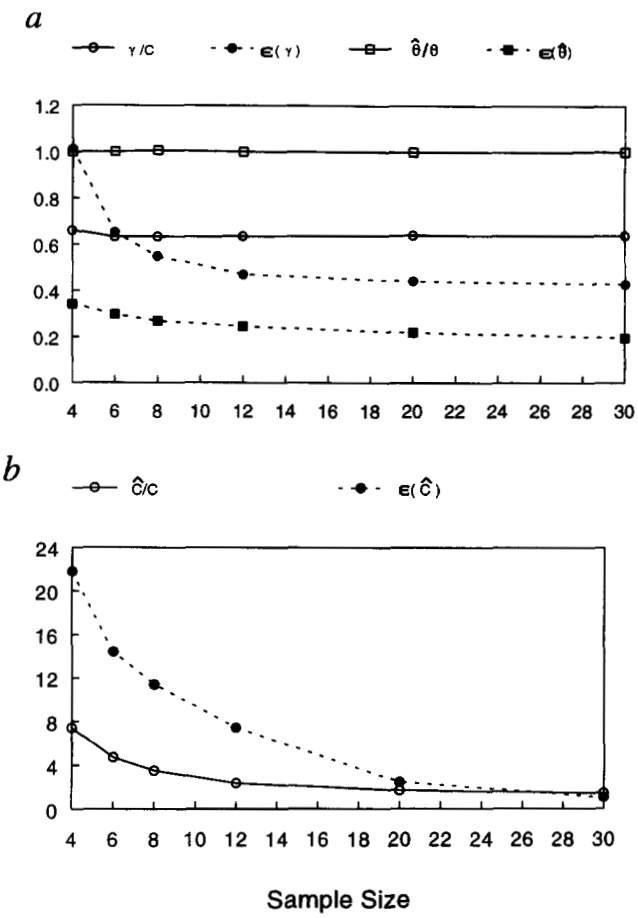

FIGURE 4.—Bias and ϵ (see text) as a function of the number of DNA sequences in the sample. Between 50,000 ($n = 4$) and 1000 ($n = 30$) independent simulations were conducted. Parametric values were $C = 0.02$, and $\theta = 0.005$. Each DNA sequence was 2000 base pairs in length. For each simulation, all possible subsets of four sequences, up to a maximum of 2000, were included in the determination of γ. For those sample sizes with more than 2000 possible four item subsets (*i.e.*, sample size >16), 2000 randomly drawn subsets were used. (a) Values for γ and WATTERSON's estimator of θ, $\hat{\theta}$ (WATTERSON 1975; HUDSON 1990). (b) Values for HUDSON's estimator (1987), $\hat{C}$.

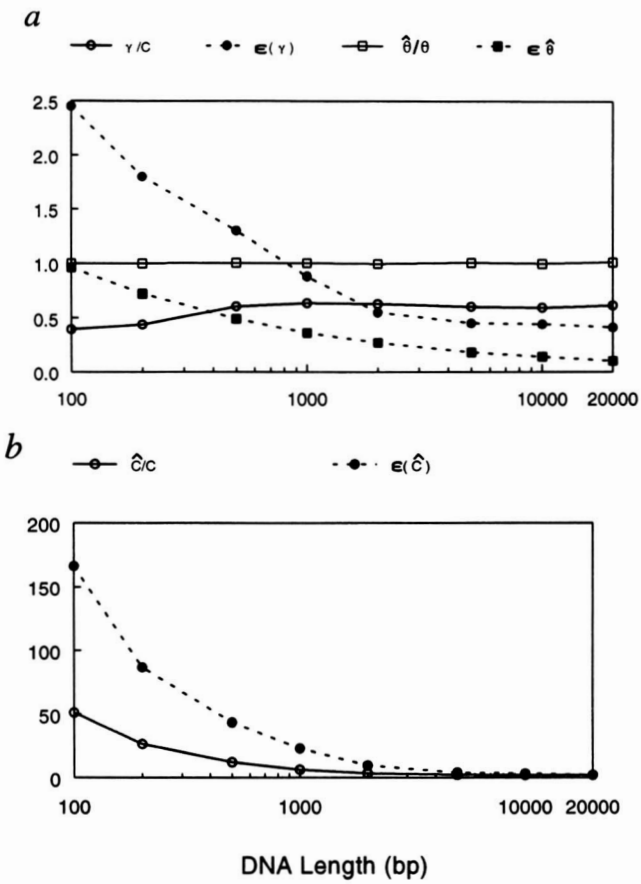

FIGURE 5.—Bias and ϵ (see text) as a function of the length of DNA sequences. Between 50,000 (100 bp) and 300 (20,000 bp) independent simulations were conducted. Parametric values were $C = 0.02$, and $\theta = 0.005$, and the sample size was 8 for all simulations. For each simulation, all possible subsets of four sequences, up to a maximum of 2000, were included in the determination of γ. For those sample sizes with >2000 possible four-item subsets (*i.e.*, sample size >16), 2000 randomly drawn subsets were used. (a) Values for γ and $\hat{\theta}$. (b) Values for HUDSON's estimator (1987), $\hat{C}$.

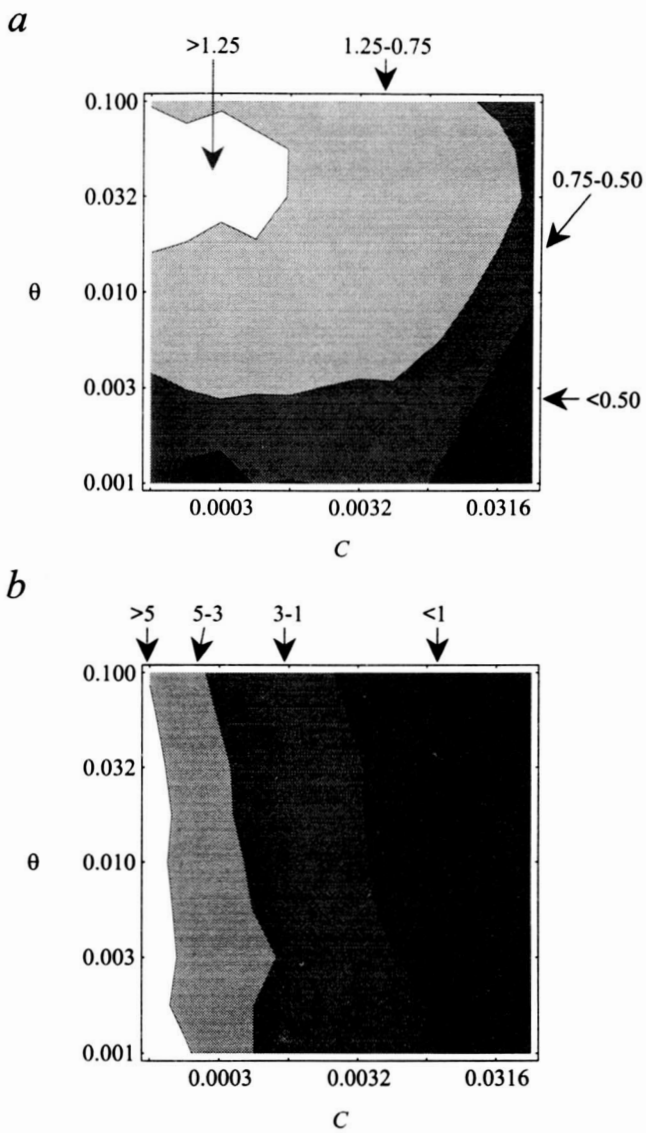

FIGURE 6.— γ bias (a) and ϵ (b) as a function of C and θ. θ goes from 0.001 to 0.1 on a log scale. C goes from 0.0001 to 0.05 on a log scale. The value is shown by arrows and indicated ranges, and the level of gray. For bias, the most extreme values were 0.23 and 1.37. For ϵ, the most extreme values were 0.33 and 7.87. Between 1000 (high values of C and θ) and 20,000 (low values of C and θ) independent simulations were conducted for each of 108 points evenly distributed throughout the log-log scale parameter space. The sample size was eight sequences, and the DNA length was 2000 bp for all simulations.

of bias and ϵ values that were observed for γ, when sample sizes are large.

Figure 5 shows bias and ϵ as a function of changing DNA length, for a sample size of eight DNA sequences. In this case, the bias of γ is most extreme for the shortest DNA sequences. Again, HUDSON's estimator has high bias and ϵ for short sequences. The simulations for Figure 5 where done with parametric values of $C = 0.02$, and $\theta = 0.005$, for a sample size of eight DNA sequences.

To examine bias and ϵ as a function of θ and C, simulations were conducted for many points across a wide plain of space for these two quantities (Figure 6). Figure 6a shows that for about half of the explored parameter space, the bias of γ is <25% (*i.e.*, between 0.75 and 1.25). The most extreme bias values found in generating Figure 6a were 0.22 and 1.37 (lower right and upper left, respectively). In general, γ is biased on the low side when θ is low and C is high, and it is biased on the high side when θ is high and C is low. For a majority of the explored parameter space, γ is biased on the low side and this bias is most extreme when θ is low and C is high. One reason for this is that the maximum likelihood expression for a pair of intervals, (35), cannot yield an estimate of $\gamma_{(i,j)k}$ if both intervals are incongruent or an incongruent interval is much shorter than a congruent interval. For much of the parameter space there will be relatively few such pairs of intervals simply because the majority of intervals are congruent. However, to the extent that pairs of intervals of these classes do occur, they reveal evidence of recombination *and* they are not included the determination

TABLE 2

Representative studies and parameter estimates

Locus	Species	Reference	No. of bp	w	γ	$\hat{\theta}$	$\gamma/\hat{\theta}$	$\hat{C}$
Adh	*D. melanogaster*	KREITMAN (1983)	2650	11	0.0125	0.0057	2.21	0.0095
white	*D. melanogaster*	KIRBY and WOLFGANG (1995)	5940	15	0.0210	0.0043	4.88	0.0035
Adh, Adh-Dup	*D. pseudoobscura*	SCHAEFFER and MILLER (1993)	3450	99	0.0654	0.0225	2.91	0.1077
β-globin	Humans	FULLERTON *et al.* (1994)	3000	60	0.0017	0.0012	1.36	0.0024
Mitochondrial Control region	Humans	VIGILANT *et al.* (1991)	682	189	0.0580	0.0439	1.32	0.0672

w is the number of DNA sequences in the study. γ is an estimate of the population recombination rate, per base pair. It was calculated using expression (37) with either all $\binom{w}{4}$ possible subsamples of four sequences, or 1000 random subsamples, whichever was fewer. $\hat{\theta}$ is Watterson's estimator of $4Nu$ (WATTERSON 1975; HUDSON 1990). $\gamma/\hat{\theta}$ is an estimate of $4Nc/4Nu$, and thus an estimate of the number of recombination events per mutation event. $\hat{C}$ is an estimate of the population recombination rate, per base pair, calculated according to HUDSON (1987).

of γ. Pairs of intervals of these types are more common if recombination rates are high and there are relatively few informative sites, that is high C and low θ.

Figure 6b shows that γ is more reliable when the population rate of recombination is high. Interestingly, ϵ depends much more strongly on C than on θ. In effect, γ works nearly as well for a given rate of recombination, regardless of how much variation there is.

The same simulations used to generate Figure 6 were also used to examine the correlation between γ and $\hat{\theta}$. The average correlation across this range of parameter values was 0.17. The highest correlations were observed when both C and θ were high, however only ~25% of the total surface was associated with correlations above 0.25, and no correlations greater than 0.5 were observed (data not shown).

Applications: The results from several analyses are shown in Table 2. In *D. melanogaster,* both the *Adh* and *white* loci reveal substantial recombination, as expected (HUDSON *et al.* 1987; KIRBY and WOLFGANG 1995). An estimate of the number of recombination events per mutation event can be obtained by dividing γ by an estimate of $4Nu$. These estimates are 2.21 and 4.88 for *Adh* and *white,* respectively (Table 2). At *Adh* there were 43 polymorphic sites (KREITMAN 1983), indicating at least 43 mutations. Under the assumptions employed in the estimation of $4Nc$, the estimated number of recombination events in the history of the *Adh* sample is 95 (2.21*43). At *Adh* and *Adh-Dup* in *D. pseudoobscura,* the estimated ratio of recombination to mutation is similar to that for *D. melanogaster.* However estimates of both $4Nc$ and $4Nu$ are higher in *D. pseudoobscura.* In a sample of human *β-globin* sequences, both γ and $\hat{\theta}$ are considerably lower than in the Drosophila examples, however there still appears to be more than one recombination per mutation event.

Table 2 also shows values for HUDSON's (1987) estimator of C. With the exception of the *white* locus estimate, the values do not differ greatly from those for γ. The data sets in Table 2 are relatively large. The *D. pseudoobscura* data in particular is of sufficient size that HUDSON's estimator could be expected to work quite well (HUDSON 1987; SCHAEFFER and MILLER 1993).

The accuracy of the estimates in Table 2 can be roughly assessed using Figure 6. If the Drosophila data set estimates for θ and C are taken as correct, then the corresponding estimates of bias and error can be obtained from the lower right portions of Figure 6, a and b, respectively. In this region the bias is near 0.5 and the standardized error ϵ is less than one. Since the bias is not a function of the sample size (Figures 4 and 5), Figure 6a should be useful as a rough guide for data sets that are larger than those used for the simulations used to generate the figure.

A critical assumption underlying the estimation of C is the infinite sites mutation model. If this does not hold, and multiple mutations have caused polymorphisms at individual base positions, then pairs of informative sites may appear to be incongruent in the absence of recombination. As an example of this, a mitochondrial data set that had been reported to have had multiple mutation events at many base positions (VIGILANT *et al.* 1991; WAKELEY 1993) was analyzed. With no recombination in the mitochondria, any pattern of incongruency in these data must be due to some failure of the infinite sites mutation model. The estimates of C are quite high (Table 2).

DISCUSSION

The estimator γ has several attractive features. Over much of the θ and C parameter space, γ has low to moderate bias and, especially for higher values of C, a low mean square error (Figure 6). Comparisons of the reliability of γ with that of a widely used estimator of θ has shown that the two parameters can be estimated with comparable reliability, especially for higher values of C. γ is also relatively independent of estimates of θ in two respects. First, an estimate of θ is not required to calculate γ so that the variance of an estimate of θ does not contribute to the variance of γ. Second, the mean square error of γ does not vary greatly as a function of θ, so that C can be estimated with similar reliability whether there are many polymorphisms or few polymorphisms.

Depending on the size of a data set, γ calculated

using expression (37) employs multiple interval-pair-based maximum likelihood estimates that are averaged across pairs of informative site intervals and subsets of four DNA sequences. The purpose of averaging across multiple pairs of intervals and across multiple subsets of four sequences is to limit the overall bias of γ to that found in estimates based on individual pairs of intervals. The averaging of many estimates, each based on a small subset of the data, should help ensure that the bias of the estimator does not change as a function of the size of a data set. It is possible to design other estimators of C based on expression (34), and we have considered several. A least squares estimator analogous to (37) has very similar properties, with slightly higher average values for the bias, γ/c (results not shown). We also considered a different maximum likelihood estimator that, rather than using the likelihood for all pairs of intervals within a sample of four sequences, employed the joint likelihood for all intervals within a subset of four sequences. This estimator also worked nearly as well as expression (37), but it had higher bias when there were very large numbers of intervals (results not shown).

Assumptions: The theory underlying γ includes important assumptions of panmixia, constant population size, infinite sites mutation model, and selective neutrality of mutations. If panmixia does not hold, and there exists population structure, then pairs of homologous DNA sequences that have the opportunity to undergo recombination will, on average, be more similar than pairs drawn at random from the entire population. Population structure will cause the genealogies that are juxtaposed by recombination to be more similar, on average, and the probability of recombination being detectable will be less. In effect, the estimate of $4Nc$ will be reduced because the rate of detectable recombination reflects smaller local effective population sizes.

The effect of changing population size on γ may be difficult to predict. On the one hand, γ is calculated using polymorphisms that are phylogenetically informative. Any recent change in the population size may be expected to affect the number of low frequency polymorphisms, caused on average by recent mutations (TAJIMA 1989a), but not so much the intermediate frequency polymorphisms used for calculating γ. On the other hand, the recombination events that are detectable by the congruency criteria are relatively new, as they have occurred more recently than the common ancestors in samples of size four. On balance, γ is expected to be fairly sensitive to the assumption of constant population size because it is more sensitive to the amount of recombination than it is to the amount of variation (Figure 6b).

If the infinite sites mutation model does not hold and multiple mutations, particularly parallel or back mutations, are segregating at individual sites, then γ will be elevated. In a phylogenetic context, the incongruency criteria is equivalent to a test for homoplasy. A pair of incongruent sites in a sample of four sequences can be explained by two events: one mutation and one recombination event (as assumed in the design of γ), *or* two mutations. These two explanations of incongruency do differ, however, in their predictions regarding flanking markers. For example, if there has just been a single recombination event, and it has caused informative sites A and B (ordered left to right) to be incongruent, then all flanking markers to the left of A should be congruent with A and incongruent with B. Similarly, sites to the right of B will be congruent with B and incongruent with A. However if A and B are incongruent because of multiple mutations, then whichever site was the target of multiple mutations will be incongruent with all other sites. This type of contrast can be used as a test for recombination and to identify the location of recombination events (STEPHENS 1985). It may be possible to develop expressions that are analogous to (24), (25), (26) and (30) that are conditioned on the probability of sites being informative because of multiple mutation events. At present, γ should only be trusted in those cases where it seems that parallel or back mutations have been rare.

The theory also assumes that mutations are neutral and that the genealogies have not been shaped by natural selection. If natural selection is stabilizing, and simply removing deleterious mutations from the population, then it may still be the case that segregating mutations are neutral and that the theory behind γ holds approximately. However, if there has been balancing selection, or recent selective sweeps then this may not be the case. Both of these kinds of selection have parallels in models of population structure. Balancing selection, whereby multiple functional alleles at a locus are maintained by selection over long periods of time can create a kind of genealogy that resembles a stable pattern of population subdivision (HEY 1991). So too, a recent selective sweep and associated genetic hitchhiking can create a pattern just like that of a recent population expansion (TAJIMA 1989b).

The theory also contains some implicit assumptions about the recombination process. Like mutation, recombination was modeled as a random process that has a low probability of occurrence per generation per pair of adjacent base pairs. In building up the expression for $I_{2,n}$ it was also implicitly assumed that every recombination event changes the tree topology, not just for adjacent base positions, but also for all of the flanking sequence. In essence recombination has been modeled as a crossover event that begins at one point and then migrates in one direction to the end of the DNA sequence. This single point process is modeled after meiotic recombination and has just one parameter, c. Other multi-point processes of gene exchange, such as gene conversion (HILLIKER *et al.* 1994) and gene exchange via induction in *Escherichia coli* (MCKANE and MILKMAN 1995), require more than two parameters to model. γ can be calculated for any data set with multi-

ple informative polymorphisms, but it is not clear how to interpret the number if recombination is actually a multiparameter process. In general, however, any process that elevates incongruency will also elevate γ, so it may still be useful as a rough indicator of gene exchange.

Estimating N and u: One surplus benefit to having a relatively reliable estimate of C is the ability in some instances to estimate N, the effective population size. For organisms with well mapped genomes it is possible to estimate c, the recombination rate per base pair per generation. For example, KLIMAN and HEY (1994) used the relationship between physical and genetic maps to generate estimates of c for many *D. melanogaster* loci. For *Adh*, the estimate of the recombination rate was 0.00198 centimorgans per kilobase pair per generation, for an estimated c of $1.98*10^{-8}$ recombination events per base pair per generation. Since γ for *Adh* was 0.0125 (Table 2), the estimate of N is 157,828 individuals (*i.e.*, $0.0125/4/1.98*10^{-8}$). Actually, this is just the estimated effective number of females, since recombination does not occur in male *D. melanogaster*.

A similar approach can be used in conjunction with an estimate of θ to estimate the neutral mutation rate, u; though the estimate will not be independent of that for N. In general, direct estimates of u are difficult to obtain. For those organisms where c is more easily estimated than u, γ provides an indirect way of estimating u. Again using *Adh* in *D. melanogaster* as an example, Table 2 shows the estimated ratio of recombination to neutral mutation rates to be 2.21. Assume, for arguments sake, that both recombination and mutation occur equally in both sexes but at half the rate observed in females. Then the *Adh* estimate of c ($1.98*10^{-8}/2$ per base pair per generation, with the one half adjustment for no recombination in males) corresponds to an estimated neutral mutation rate of $4.5*10^{-9}$ per base pair per generation ($((1.98*10^{-8}/2)*(1/2.21))$).

This work was supported by National Science Foundation grant DEB-9306625 to J.H. and National Institutes of Health National Research Service Award GM-17745-01 to J.W.

LITERATURE CITED

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer Verlag, New York.

FULLERTON, S. M., R. M. HARDING, A. J. BOYCE and J. B. CLEGG, 1994 Molecular and population genetic analysis of allelic sequence diversity at the human β-globin locus. Proc. Natl. Acad. Sci. USA **91:** 1805–1809.

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

HEY, J., 1991 A multi-dimensional coalescent process applied to multi-allelic selection models and migration models. Theor. Pop. Biol. **39:** 30–48.

HILLIKER, A. J., G. HARAUZ, A. G. REAUME, M. GRAY, S. H. CLARK *et al.*, 1994 Meiotic gene conversion tract length distribution within the *rosy* locus of *Drosophila melanogaster*. Genetics **137:** 1019–1026.

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

HUDSON, R. R., 1983b Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. Camb. **50:** 245–250.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by P. H. HARVEY and L. PARTRIDGE. Oxford University Press, New York.

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics **61:** 893–903.

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

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

KIRBY, D. A., and S. WOLFGANG, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

KLIMAN, R. M., and J. HEY, 1994 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

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

KUHNER, M. K., J. YAMATO and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics **140:** 1421–1430.

MCKANE, M., and R. MILKMAN, 1995 Transduction, restriction and recombination patterns in *Escherichia coli*. Genetics **139:** 35–43.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotide sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

STEPHENS, J. C., 1985 Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. Mol. Biol. Evol. **2:** 539–556.

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

TAJIMA, F., 1989a The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

TAJIMA, F., 1989b Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

TAJIMA, F., 1993 Measurement of DNA polymorphism, pp. 37–59 in *Mechanisms of Molecular Evolution*, edited by N. TAKAHATA and A. G. CLARKE. Sinauer Associates, Sunderland, MA.

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES and A. C. WILSON, 1991 African populations and the evolution of human mitochondrial DNA. Science **253:** 1503–1507.

WAKELEY, J., 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:** 613–623.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. **7:** 256–275.

Communicating editor: R. R. HUDSON

APPENDIX

All possible histories of two adjacent sites for cases I–IV

Case	Genealogy	Probability	$E(u\tau_a)$	$E(u\tau_b)$	$E(u^2\tau_a\tau_b)$
I	(AB:A-,-B)(A-,-B:1)(1,AB:2)(2,AB:3)(3,AB:4)	$\frac{1}{15}$	$\frac{\theta}{6}$	$\frac{\theta}{6}$	$\frac{\theta^2}{18}$
	(AB:A-,-B)(A-,-B:1)(1,AB:2)(AB,AB:3)(2,3:4)	$\frac{1}{30}$	$\frac{7\theta}{6}$	$\frac{7\theta}{6}$	$\frac{43\theta^2}{18}$
	(AB:A-,-B)(AB,AB:1)(A-,-B:2)(1,2:3)(3,AB:4)	$\frac{1}{60}$	$\frac{\theta}{4}$	$\frac{\theta}{4}$	$\frac{7\theta^2}{72}$
	(AB:A-,-B)(AB,AB:1)(A-,-B:2)(1,AB:3)(2,3:4)	$\frac{1}{60}$	$\frac{\theta}{4}$	$\frac{\theta}{4}$	$\frac{7\theta^2}{72}$
	(AB:A-,-B)(AB,AB:1)(A-,-B:2)(2,AB:3)(2,3:4)	$\frac{1}{60}$	$\frac{5\theta}{4}$	$\frac{5\theta}{4}$	$\frac{187\theta^2}{72}$
	(AB:A-,-B)(AB,AB:1)(1,AB:2)(A-,-B:3)(2,3:4)	$\frac{1}{60}$	$\frac{\theta}{12}$	$\frac{\theta}{12}$	$\frac{\theta^2}{72}$
	(AB:A-,-B)(AB,AB:1)(1,AB:2)(2,-B:3)(3,A-:4)	$\frac{1}{30}$	$\frac{\theta}{12}$	$\frac{\theta}{12}$	$\frac{\theta^2}{72}$
	(AB:A-,-B)(AB,AB:1)(1,A-:2)(2,-B:3)(3,AB:4)	$\frac{1}{30}$	$\frac{\theta}{12}$	$\frac{\theta}{4}$	$\frac{\theta^2}{36}$
	(AB:A-,-B)(AB,AB:1)(1,A-:2)(AB,-B:3)(2,3:4)	$\frac{1}{30}$	$\frac{\theta}{12}$	$\frac{5\theta}{4}$	$\frac{\theta^2}{9}$
	(AB:A-,-B)(AB,AB:1)(1,-A:2)(2,AB:3)(3,-B:4)	$\frac{1}{30}$	$\frac{\theta}{12}$	$\frac{\theta}{4}$	$\frac{\theta^2}{36}$
	(AB:A-,-B)(AB,AB:1)(-B,AB:2)(1,A-:3)(2,3:4)	$\frac{1}{30}$	$\frac{\theta}{4}$	$\frac{17\theta}{12}$	$\frac{10\theta^2}{24}$
	(AB:A-,-B)(AB,AB:1)(-B,AB:2)(2,A-:3)(1,3:4)	$\frac{1}{30}$	$\frac{5\theta}{4}$	$\frac{17\theta}{12}$	$\frac{17\theta^2}{6}$
	(AB:A-,-B)(AB,AB:1)(-B,AB:2)(1,2:3)(3,A-:4)	$\frac{1}{30}$	$\frac{\theta}{4}$	$\frac{5\theta}{12}$	$\frac{\theta^2}{6}$
	(AB:A-,-B)(-B,AB:1)(1,A-:2)(2,AB:3)(3,AB:4)	$\frac{1}{15}$	$\frac{\theta}{6}$	$\frac{\theta}{4}$	$\frac{5\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(1,A-:2)(AB,AB:3)(2,3:4)	$\frac{1}{30}$	$\frac{5\theta}{4}$	$\frac{7\theta}{6}$	$\frac{179\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(AB,AB:2)(1,A-:3)(2,3:4)	$\frac{1}{30}$	$\frac{7\theta}{6}$	$\frac{17\theta}{12}$	$\frac{65\theta^2}{24}$
	(AB:A-,-B)(-B,AB:1)(AB,AB:2)(1,2:3)(3,A-:4)	$\frac{1}{30}$	$\frac{\theta}{6}$	$\frac{5\theta}{12}$	$\frac{\theta^2}{8}$
	(AB:A-,-B)(-B,AB:1)(AB,AB:2)(2,A-:3)(1,3:4)	$\frac{1}{30}$	$\frac{\theta}{6}$	$\frac{17\theta}{12}$	$\frac{7\theta^2}{24}$
	(AB:A-,-B)(-B,AB:1)(1,AB:2)(2,A-:3)(3,AB:4)	$\frac{1}{15}$	$\frac{\theta}{12}$	$\frac{\theta}{6}$	$\frac{\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(1,AB:2)(2,AB:3)(3,A-:4)	$\frac{1}{15}$	$\frac{\theta}{12}$	$\frac{\theta}{6}$	$\frac{\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(1,AB:2)(AB,A-:3)(2,3:4)	$\frac{1}{15}$	$\frac{\theta}{12}$	$\frac{7\theta}{6}$	$\frac{7\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(AB,A-:2)(1,2:3)(3,AB:4)	$\frac{1}{15}$	$\frac{\theta}{4}$	$\frac{\theta}{6}$	$\frac{5\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(AB,A-:2)(1,AB:3)(2,3:4)	$\frac{1}{15}$	$\frac{\theta}{4}$	$\frac{7\theta}{6}$	$\frac{23\theta^2}{72}$
	(AB:A-,-B)(-B,AB:1)(AB,A-:2)(2,AB:3)(1,3:4)	$\frac{1}{15}$	$\frac{\theta}{6}$	$\frac{5\theta}{4}$	$\frac{17\theta^2}{72}$
II	(AB,AB:1)(1:A-,-B)(A-,-B:2)(2,AB:3)(3,AB:4)	$\frac{1}{27}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(7C^2+40C+68)}{72(C+2)^2}$
	(AB,AB:1)(1:A-,-B)(A-,-B:2)(AB,AB:3)(2,3:4)	$\frac{1}{54}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta^2(187C^2+808C+884)}{72(C+2)^2}$

APPENDIX

Continued

Case	Genealogy	Probability	$E(u\tau_a)$	$E(u\tau_b)$	$E(u^2\tau_a\tau_b)$
II	(AB,AB:1)(1:A-,-B)(AB,AB:2)(2,-B:3)(3,A-:4)	$\frac{1}{27}$	$\frac{\theta(17C+38)}{12(C+2)}$	$\frac{\theta(5C+14)}{12(C+2)}$	$\frac{\theta^2(51C^2+248C+308)}{72(C+2)^2}$
	(AB,AB:1)(1:1-,-B)(AB,AB:2)(A-,-B:3)(2,3:4)	$\frac{1}{54}$	$\frac{\theta(17C+38)}{12(C+2)}$	$\frac{\theta(17C+38)}{12(C+2)}$	$\frac{\theta^2(255C^2+968C+1052)}{72(C+2)^2}$
	(AB,AB:1)(1:A-,-B)(AB,-B:2)(2,A-:3)(3,AB:4)	$\frac{2}{27}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+20)}{36(C+2)^2}$
	(AB,AB:1)(1:A-,-B)(AB,-B:2)(2,AB:3)(3,A-:4)	$\frac{2}{27}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+14)}{9(C+2)^2}$
	(AB,AB:1)(1:A-,-B)(AB,-B:2)(A-,AB:3)(2,3:4)	$\frac{2}{27}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+20)}{36(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,-B:2)(1,2:3)(3,AB:4)	$\frac{1}{27}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(7C^2+40C+68)}{72(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,-B:2)(2,AB:3)(1,3:4)	$\frac{1}{27}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta^2(187C^2+808C+884)}{72(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,-B:2)(1,AB:3)(2,3:4)	$\frac{1}{27}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(7C^2+40C+68)}{72(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(1,A-:2)(2,-B:3)(3,AB:4)	$\frac{2}{27}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+20)}{36(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(1,A-:2)(AB,-B:3)(2,3:4)	$\frac{2}{27}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+14)}{9(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(1,A-:2)(2,AB:3)(3,-B:4)	$\frac{2}{27}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+20)}{36(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(1,AB:2)(2,A-:3)(3,-B:4)	$\frac{2}{27}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+28)}{72(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(1,AB:2)(A-,-B:3)(2,3:4)	$\frac{1}{27}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta(C+6)}{12(C+2)}$	$\frac{\theta^2(C^2+8C+28)}{72(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,AB:2)(1,-B:3)(2,3:4)	$\frac{2}{27}$	$\frac{\theta(17C+38)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(15C^2+80C+108)}{36(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,AB:2)(2,-B:3)(1,3:4)	$\frac{2}{27}$	$\frac{\theta(17C+38)}{12(C+2)}$	$\frac{\theta(15C+34)}{12(C+2)}$	$\frac{\theta^2(51C^2+220C+240)}{18(C+2)^2}$
	(AB,AB:1)(AB:A-,-B)(A-,AB:2)(1,2:3)(3,-B:4)	$\frac{2}{27}$	$\frac{\theta(5C+14)}{12(C+2)}$	$\frac{\theta(3C+10)}{12(C+2)}$	$\frac{\theta^2(3C^2+16C+24)}{18(C+2)^2}$
III	(AB,AB:1)(1,AB:2)(2:A-,-B)(A-,-B:3)(3,AB:4)	$\frac{2}{9}$	$\frac{\theta}{3(C+2)}$	$\frac{\theta}{3(C+2)}$	$\frac{2\theta^2}{9(C+2)^2}$
	(AB,AB:1)(1,AB:2)(2:A-,-B)(AB,-B:3)(3,A-:4)	$\frac{4}{9}$	$\frac{\theta}{3(C+2)}$	$\frac{\theta}{3(C+2)}$	$\frac{2\theta^2}{9(C+2)^2}$
	(AB,AB:1)(AB,AB:2)(1:A-,-B)(A-,-B:3)(2,3:4)	$\frac{1}{9}$	$\frac{\theta(2C+3)(2C+5)}{3(C+1)(C+2)}$	$\frac{\theta(2C+3)(2C+5)}{3(C+1)(C+2)}$	$\frac{2\theta^2(13C^4+94C^3+258C^2+319C+151)}{9(C+1)^2(C+2)^2}$
	(AB,AB:1)(AB,AB:2)(1:A-,-B)(2,-B:3)(3,A-:4)	$\frac{2}{9}$	$\frac{\theta(2C+3)(2C+5)}{3(C+1)(C+2)}$	$\frac{\theta(C^2+7C+9)}{3(C+1)(C+2)}$	$\frac{\theta^2(5C^4+50C^3+186C^2+299C+176)}{9(C+1)^2(C+2)^2}$
IV	(AB,AB:1)(1,AB:2)(2,AB:3)	$\frac{2}{3}$	$\frac{\theta}{3(C+2)}$	$\frac{\theta}{3(C+2)}$	$\frac{2\theta^2}{9(C+2)^2}$
	(AB,AB:1)(AB,AB:2)(1,2:3)	$\frac{1}{3}$	$\frac{\theta(4C+7)}{3(C+1)(C+2)}$	$\frac{\theta(4C+7)}{3(C+1)(C+2)}$	$\frac{2\theta^2(13C^2+47C+43)}{9(C+1)^2(C+2)^2}$

Every genealogy above of two sites in a sample of four sequences is defined by either three coalescent events (case IV) or one recombination event and four coalescent events (cases I–III). The parenthetical notation lists these events from most recent to most ancient, left to right. Inside each set of parentheses, descendant(s) are listed to the left of the colon and ancestor(s) to the right. Thus, in case I, genealogy 1, the first event looking back is a recombination event; the second is the coalescence of the resulting partial ancestral sequences; the third is the coalescence of the result of the second event with any one of the other three other sampled sequences; the fourth is the coalescence of the result of the third event with either of the remaining two sequences; and the fifth is the only possible event left. Certain symmetry properties are used to avoid enumerating trees that are redundant with respect to the derivation of I_0, I_1, and I_2. For instance, in case I, genealogy 1, it does not matter which of the four sequences is the result of recombination (first event), neither does it matter in the third and fourth events, which of the presently sampled sequences, is chosen to coalesce.