The Ancestry of a Sample of Sequences Subject to Recombination

Carsten Wiuf and Jotun Hein

Institute of Biological Sciences, University of Aarhus, DK-8000 Aarhus, Denmark

Manuscript received February 10, 1998
Accepted for publication November 30, 1998

ABSTRACT

In this article we discuss the ancestry of sequences sampled from the coalescent with recombination with constant population size $2N$. We have studied a number of variables based on simulations of sample histories, and some analytical results are derived. Consider the leftmost nucleotide in the sequences. We show that the number of nucleotides sharing a most recent common ancestor (MRCA) with the leftmost nucleotide is $\approx\log(1 + 4NLr)/4Nr$ when two sequences are compared, where L denotes sequence length in nucleotides, and r the recombination rate between any two neighboring nucleotides per generation. For larger samples, the number of nucleotides sharing MRCA with the leftmost nucleotide decreases and becomes almost independent of $4NLr$. Further, we show that a segment of the sequences sharing a MRCA consists in mean of $3/8Nr$ nucleotides, when two sequences are compared, and that this decreases toward $1/4Nr$ nucleotides when the whole population is sampled. A measure of the correlation between the genealogies of two nucleotides on two sequences is introduced. We show analytically that even when the nucleotides are separated by a large genetic distance, but share MRCA, the genealogies will show only little correlation. This is surprising, because the time until the two nucleotides shared MRCA is reciprocal to the genetic distance. Using simulations, the mean time until all positions in the sample have found a MRCA increases logarithmically with increasing sequence length and is considerably lower than a theoretically predicted upper bound. On the basis of simulations, it turns out that important properties of the coalescent with recombinations of the whole population are reflected in the properties of a sample of low size.

UNDERSTANDING the genealogical relationship between sequences in a diploid population has been central to recent analyses of the dynamics of sequence evolution at the population level. The stochastic process generating the genealogical relationship between k sampled sequences from a population with constant size N and no recombination was first described by Watterson (1975) and further developed into the theory of the coalescent by Kingman (1982). The process of evolution of sequences subject to both coalescence and recombination in a population was first described by Hudson (1983). In Hudson's approach the combined coalescent and recombination process is followed back in time until any position in the extant sequences has found a most recent common ancestor (MRCA). Distant positions will not necessarily share the same history, and the ancestral positions can be located on different sequences. However, the genealogies of distinct but linked positions are correlated: Positions far apart have ancestries almost independent of each other, whereas positions close to each other tend to have identical ancestry. Griffiths and Marjoram (1997) proved that the set of MRCAs to a sample is finite for any sample size. But, even positions sharing the same MRCA can have very different histories.

In this article we discuss the ancestry of a sample of k sequences subject to both coalescence and recombination. This is done mainly through simulations of sample histories. The combinatorial complexity of the coalescent with recombination makes exact results difficult to derive and, in most cases, restricted to samples of size 2. We measure the sequence length in expected number of recombinations per sequence per $2N$ generations, where N is population size. The population size is assumed to be constant from generation to generation.

Our results can be broadly divided into two parts. In the first part, we have focused on the structure of a single MRCA. Consider the MRCA at position 0 of the sequences. Call this ancestor $\text{MRCA}_k(0)$, where k refers to sample size. If there is no recombination in the history of the sample, all positions $q > 0$ will share the same MRCA, *i.e.*, $\text{MRCA}_k(0) = \text{MRCA}_k(q)$. However, if recombination is present, only a subset of the positions $q > 0$ will share this MRCA. In the example in Figure 1 positions $0 \le q < \frac{1}{4}$ and $\frac{1}{2} \le q \le 1$ share MRCA spread on two distinct segments, while $\frac{1}{4} \le q < \frac{1}{2}$ share MRCA.

Furthermore, we are interested in the following variables: (1) length of ancestral material that shares MRCA with position 0 [in the example in Figure 1 this amounts to $(1 - \frac{1}{2}) + (\frac{1}{4} - 0) = \frac{3}{4}$], (2) the number of segments into which the positions sharing MRCA with position 0

Corresponding author: Carsten Wiuf, Department of Statistics, University of Oxford, 1 South Parks Rd., Oxford, OX1 3TG, England. E-mail: wiuf@stats.ox.ac.uk

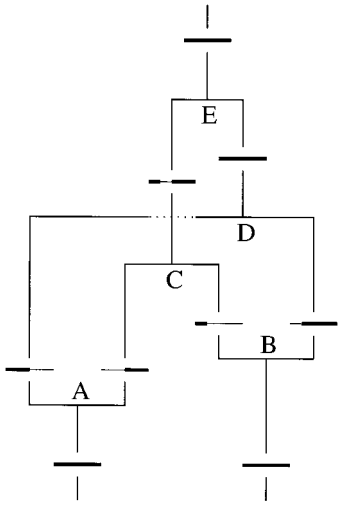

Figure 1.—The coalescent with recombination. The genealogy of a sample of size 2 is shown. Sequence length is 1 (= $\rho/2$). Time starts at present (bottom) and increases going backward in time (top). When a branch splits in two, a recombination event happens and when two branches merge, a coalescent event happens. Thick lines, material ancestral to the sample; thin lines, nonancestral material. (A) The first event going backward in time is a recombination event with breakpoint $q = \frac{1}{2}$, whereby the ancestral material to the sample is located on three sequences. (B) The second event is a recombination event with breakpoint $q = \frac{1}{4}$, spreading the ancestral material on four sequences. (C) The third event is a coalescent event creating a new sequence, say *S*. The ancestral material on *S* is partitioned into two segments with nonancestral material in between (of length $\frac{1}{4}$). This nonancestral material is trapped between the two segments of ancestral material, hence called trapped material. If a recombination event happens within the trapped material, the two segments of ancestral material are spread on two different ancestors and recombination events happening within trapped material affect the genealogy of the sample. The positions $\frac{1}{4} \leq q < \frac{1}{2}$ find a MRCA at event D, and positions $0 \leq q < \frac{1}{4}$ and $\frac{1}{2} \leq q \leq 1$ at event E.

are partitioned (in the example in Figure 1 this is two segments).

In the second part, we have focused on the time back to MRCAs. The total branch length, $G_k(q)$, and the height, $T_k(q)$ (time until a MRCA), of the genealogy of a single nucleotide are distributed according to the coalescent without recombination. In contrast, the distribution of the time until all nucleotides have found a MRCA, $T_k = \max\{T_k(q)|q\}$ and the distribution of $G_k = \max\{G_k(q)|q\}$ depend on the total genetic length ρ. We have investigated the expected values of these two variables. Moreover, we discuss a notion of shared sequence ancestry that relates to the correlation between genealogies.

The recombination rate $\rho = 4NLr$ has been varied from 0 to 50 in the simulations. The quantity r is the probability of a recombination event between any two neighboring positions in a sequence per generation, and L is number of nucleotides in a sequence. Let us assume that r is 10^{-7} in the human genome and that the effective population size of humans is 10^4. With $\rho = 50$ we have $L = \rho/(4Nr) = 50 \cdot 10^7/(2 \cdot 10^4) = 2.5 \cdot 10^4$ nucleotides. Thus, our simulation results cover the ancestry of a sample of human DNA sequences of length up to 25,000 nucleotides.

THE COALESCENT WITH RECOMBINATION

The model of a population of sequences subject to recombination is the following: Each sequence is L nucleotides long and recombination is assumed to occur to the right of a nucleotide. The population is of constant size N and diploid, *i.e.*, there are $2N$ sequences in the population.

A new generation is obtained from the present by (1) selecting with probability $1 - r$ a single parent uniformly at random and (2) selecting with probability r two parents uniformly at random and recombining these. Each sequence in the next generation chooses one or two parents in this manner. The collection of these offspring forms the next generation. The process starts at the present and time increases as it goes backward.

This process is transformed into one of a continuous time and continuous sequence by letting $N \to \infty$ and measuring time in $2N$ generations and by letting $L \to \infty$ and $r \to 0$, such that $4rLN \to \rho$. Here $2rLN$ is the expected number of recombinations per $2N$ sequences per generation. Sequence length is measured in expected number of recombinations per $2N$ sequences per generation; that is, the entire sequence length is $\rho/2$. Hudson (1983) showed that the waiting time until a sequence is created by a recombination event from two sequences is exponentially distributed with intensity parameter $\rho_0/2$. For the extant sequences, $\rho_0/2$ is simply the length of the sequences, *i.e.*, $\rho_0 = \rho$. For ancestral sequences, $\rho_0/2$ is the length of the interval spanned by regions that have ancestral material. Note that this interval *can* include regions with nonancestral material (*cf.* Figure 1). The recombination breakpoint is uniformly distributed within this material. The waiting time going backward in time until k sequences have only $k - 1$ ancestors in the population is exponentially distributed with intensity parameter $k(k - 1)/2$, and the two sequences that have a common ancestor at that time are uniformly distributed among different pairs. This was first realized by Watterson (1975), and later developed into the theory of the coalescent by Kingman (1982).

The coalescent with recombination has further been investigated by Hudson and Kaplan (1985), Kaplan and Hudson (1985), Griffiths and Marjoram (1996, 1997), and Wiuf and Hein (1997, 1999).

The genealogy of a sample of sequences can be simulated by going back in time, waiting for what occurs first, a recombination or a coalescence, and then performing the appropriate operation on the set of ancestral sequences. Recombination increases the number of

sequences carrying ancestral material by one, but does not increase the total amount of ancestral material. A coalescence decreases the number of sequences with ancestral material by one. It can increase the amount of material where recombination can occur, because coalescence can trap some nonancestral material (Figure 1). When any position on the extant sequences has found a MRCA, not necessarily the same ancestor, all segments with ancestral material spliced together constitute one sequence. Above this point, coalescence cannot reduce the amount of ancestral material and all that occurs is redistribution of ancestral material on different sequences by recombination and coalescence. Because the rate of coalescence is quadratic in the number of sequences, and the rate of recombination is at most linear, all positions eventually find a MRCA.

RESULTS

In this section we present simulated and mathematical results related to the MRCAs of a sample on k sequences (sections 1–6). We used an algorithm described in Wiuf and Hein (1999) to simulate sample histories. For each value of $k = 2$, 3, 5, 10, 25, 50, and 100 we simulated 2000 sample histories with recombination rate $\rho = 50$.

1. Definitions: We define a number of mathematical quantities that relate to the coalescent with recombination and to the results that we derive and discuss below.

Assume a sample of size k is given, with k possibly infinite. Let MRCA$_k(q)$ denote the MRCA to position q in the sample of k sequences. The time until the MRCA$_k(q)$ is distributed according to the coalescent process without recombination because one position cannot be subject to recombination. Further, let $A_k(q) = 1$ if there is a shift from one MRCA to another MRCA in position q, and $A_k(q) = 0$ otherwise; and let $B_k(q) = 1$ if there is a recombination breakpoint in position q within ancestral material, and $B_k(q) = 0$ otherwise. A_k stands for ancestor and B_k for breakpoint. We have $A_k(q) = 1$ iff the MRCAs to the left and to the right of position q are different, *i.e.*, if MRCA$_k(q - \varepsilon) \neq$ MRCA$_k(q + \varepsilon)$, provided ε is small. The definitions are illustrated in Figure 2. Both quantities $A_k(q)$ and $B_k(q)$ depend on the ancestral history of positions local to q only and not on the entire sequence history. Note that if $A_k(q) = 1$ then also $B_k(q) = 1$, but not necessarily the other way around. All recombination events do not necessarily result in a shift from one MRCA to another MRCA. Moreover, the distributions of the $A_k(q)$'s and $B_k(q)$'s, $q \geq 0$ are invariant under translations along the sequences. As an example, $(A_k(0), A_k(q))$ is distributed like $(A_k(p), A_k(q + p))$; the distribution depends on the relative distance between positions only ($q = q - 0 = q + p - p$) and not on the actual positions. This makes the two processes stationary processes (Daley and Vere-Jones 1988).

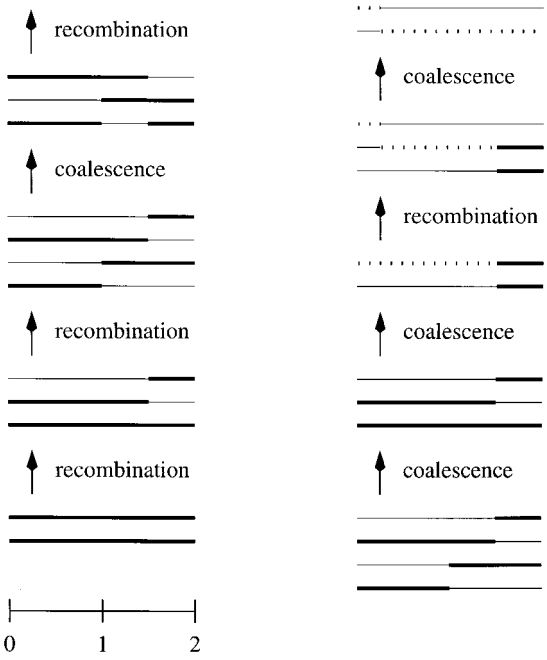

Figure 2.—An illustration of the definitions of $B_k(q)$, $A_k(q)$, $R_k(\rho)$, and r_k. The figure shows the ancestral history of a sample of size 2 until all positions have found a MRCA. Thick lines, material ancestral to the sample; thin lines, nonancestral material. When a position has found a MRCA, it is marked with a dot. The recombination rate is $\rho = 4$, so sequence length is 2. The first two events are recombinations, spreading the ancestral material on four sequences. The fourth event is also a recombination, but in nonancestral (trapped) material. The sixth event is a coalescence whereby positions $0 \leq q < 1.5$ find a MRCA. The seventh event is a recombination in ancestral material but after the position has found a MRCA. Finally, the positions $1.5 \leq q < 2$ find a MRCA. In total we find that the number of recombination events within ancestral material, but before the positions find a MRCA, is $R_2(4) = 2$, and the positions where this happens fulfill $B_2(1) = B_2(1.5) = 1$. For all other positions $B_2(q) = 0$. The length until the first recombination breakpoint counted from position 0 is 1, *i.e.*, $r_2 = 1$. In total there are three ancestral sequences where positions find MRCA, events five, six, and eight. The shifts from one MRCA to another MRCA happen in position 1.5 only, so that $A_2(1.5) = 1$ and $A_2(q) = 0$ for $q \neq 1.5$. Thus, we have $S_2(4) = 2$.

On the basis of the definitions of $A_k(q)$ and $B_k(q)$ we define

$$R_k(\rho) = \#\{B_k(q) = 1\,; \quad 0 \leq q \leq \rho/2\}, \tag{1}$$

$$S_k(\rho) = \#\{A_k(q) = 1\,; \quad 0 \leq q \leq \rho/2\} + 1. \tag{2}$$

The variable $R_k(\rho)$ is the number of recombination events within ancestral material until all positions have found a MRCA (*cf.* Figure 2). Note that this is not necessarily the same MRCA for all positions. Similarly, $S_k(\rho)$ is the number of shifts from one MRCA to another MRCA plus one. The material sharing a MRCA is partitioned into disjoint and distinct segments (as illustrated in Figure 3). The total number of segments equals the variable $S_k(\rho)$. Trapped material between two such seg-

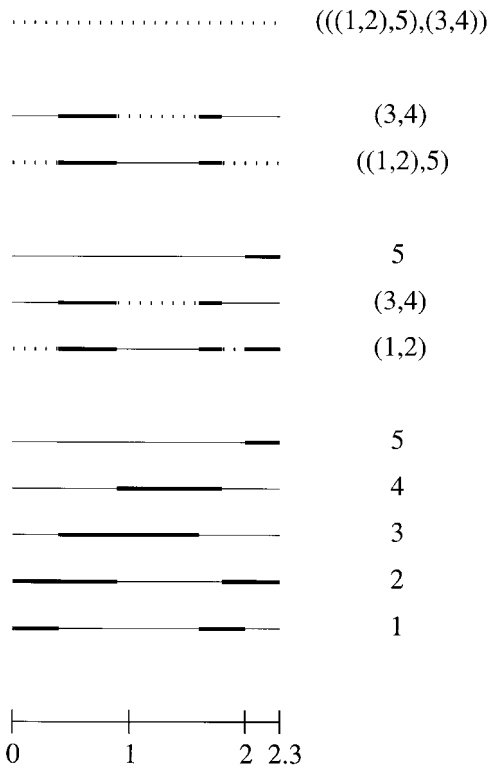

Figure 3.—An illustration of the definitions of $S_k(\rho)$ and s_k. The figure shows the ancestral history of a sample of size 2 until all positions have found a MRCA. Thick lines, material ancestral to the sample; thin lines, nonancestral material. When a position has found a MRCA, it is marked with a dot. The recombination rate is $\rho = 4.6$, so sequence length is 2.3. The first five events are recombination events within ancestral material spreading the ancestral material on 7 sequences (not shown). Then two coalescence events happen joining parts of the ancestral material (shown in the figure). After this, sequence 1 and 2 coalesce whereby positions $0 \le q$ 0.4 and $1.8 \le q < 2.0$ find a MRCA [the sequence marked (1,2)], and sequence 3 and 4 coalesce whereby positions $0.9 \le q < 1.6$ find a MRCA [the sequence marked (3,4)]. The next event joins sequence 5 with (1,2), called ((1,2),5), and finally sequences ((1,2),5) and (3,4) coalesce into (((1,2),5),(3,4)). The leftmost MRCA, or the MRCA to position 0 in the sample is sequence (1,2). The amount of material sharing MRCA with position 0 is $L_2(4.6) = (2.0 - 1.8) + (0.4 - 0) = 0.6$. The leftmost MRCA consists of two segments, $0 \le q < 0.4$ and $1.8 \le q < 2.0$, that share MRCA with position 0. Further, there are two trapped segments ancestral to the sample, $0.4 \le q < 0.9$ and $1.6 \le q < 1.8$, and one trapped nonancestral segment, $0.9 \le q < 1.6$. Sequence (3,4) is the MRCA to positions $0.9 \le q < 1.6$, sequence ((1,2),5) to $2.0 \le q < 2.3$, and sequence (((1,2),5),(3,4)) to $0.4 \le q < 0.9$ and $1.6 \le q < 1.8$. In total there are five shifts from one MRCA to another MRCA, *i.e.*, $S_2(4.6) = 5 + 1 = 6$, and the length until the first shift counted from position 0 is 0.4.

ments on the same sequence can either be ancestral to the sample or nonancestral (Figure 3). In the former case, a position within the trapped material has not yet found a MRCA.

The number $R_k(\rho)$ was first studied by Hudson and Kaplan (1985), and $S_k(\rho)$ by Griffiths and Marjoram (1997). We call $S_k(\rho)$ the number of segments carrying ancestral material in the set of MRCAs. In light of the above discussion, this can be slightly misleading but is kept for matters of convenience.

In what follows we denote the length of a sequence by $R \equiv \rho/2$. Let

$$L_k(\rho) = \lim_{\varepsilon \to 0} \varepsilon \sum_{i=1}^{R\varepsilon^{-1}} 1\{\text{MRCA}_k(0) = \text{MRCA}_k(i\varepsilon)\}, \tag{3}$$

where $1\{\cdot\}$ denotes the indicator function of a set. This function takes the value 1 if the condition in the bracket is fulfilled and zero otherwise. The variable $L_k(\rho)$ measures the amount of positions sharing MRCA with position 0. This amount is (potentially) just a subset of the entire material on $\text{MRCA}_k(0)$ ancestral to the sample: As illustrated in Figure 3, there can be ancestral material on $\text{MRCA}_k(0)$ that does not share MRCA with position 0.

If the coalescent process with recombination is studied on a grid of points equally spaced with distance ε (in contrast to a continuum of points), the definition of $L_k(\rho)$ would be $L_k(\rho) = \varepsilon \sum_{i=1}^{R\varepsilon^{-1}} 1\{\text{MRCA}_k(0) = \text{MRCA}_k(i\varepsilon)\}$; that is, the number of times the $\text{MRCA}_k(0)$ is visited moving along the sequences multiplied by the distance between the points.

We call $\text{MRCA}_k(0)$ (or the MRCA to position 0), the leftmost MRCA, and $L_k(\rho)$ the amount of material sharing MRCA with position 0.

2. Segment length: Hudson and Kaplan (1985) showed that the number, $R_k(\rho)$, of recombination events within ancestral material until all positions have found a MRCA has expectation

$$E[R_k(\rho)] = \rho \sum_{i=1}^{k-1} \frac{1}{i}. \tag{4}$$

Moreover, Griffiths and Marjoram (1997) proved that the expectation of the number, $S_k(\rho)$, of segments carrying material ancestral to the sample in the set of MRCAs is

$$E[S_k(\rho)] = 1 + \rho \left(1 - \frac{2}{k(k+1)}\right). \tag{5}$$

We are interested in the sequence length between successive recombination breakpoints and the length between successive shifts between MRCAs. The above equations give us the expected number of each kind, recombination events/breakpoints and shifts between MRCAs, in sequences of length $R \equiv \rho/2$.

Denote by r_k the length between $q = 0$ and the first recombination point along the sequences, and by s_k the length to the first shift from one MRCA to another MRCA measured from $q = 0$ (Figures 2 and 3). We here assume that sequences are potentially infinite so that there always is a first recombination event and a shift between MRCAs. Because of the stationary property of the process, it follows that the expected value of r_k, given

a recombination event happened in position $q = 0$, is

$$E[r_k|B_k(0) = 1] = \frac{1}{2\sum_{i=1}^{k-1} 1/i} \quad (6)$$

(see appendix).

Similarly, one obtains the expected value of s_k, given that there is a shift in position $q = 0$ from one MRCA to another MRCA, by

$$E[s_k|A_k(0) = 1] = \frac{1}{2 - 4/k(k+1)} \quad (7)$$

(see appendix).

The two expressions (6) and (7) hold for $k = \infty$ as well, yielding

$$E[r_\infty|B_\infty(0) = 1] = 0 \quad \text{and} \quad E[s_\infty|A_\infty(0) = 1] = \tfrac{1}{2} \quad (8)$$

(see appendix). Hence the length r_∞ between two recombination points in the ancestral material is 0 with probability one. Further, Equation 8 means that almost all recombination events are invisible in MRCAs even in large samples. From the fact that $E[S_k(\rho)] \ll E[R_k(\rho)]$ for large k, this is expected.

In Wiuf and Hein (1999) the expected length of r_k is calculated

$$E[r_k] = \frac{1}{2}\sum_{i=1}^{k-2}(-1)^{i-1}\frac{(k-1)!}{i!\,(k-i-2)!}\log(i+1) \quad (9)$$

for $2 < k < \infty$, and $E[r_2] = \infty$ and $E[r_\infty] = 0$ (see appendix). The expectation of r_k decreases in k toward 0. Griffiths and Marjoram (1996) showed that the time until a MRCA in position 0 given a recombination ($B_k(0) = 1$) is

$$2\left(1 - \frac{1}{k}\right) + \frac{2(1/k + \Sigma_{j=2}^{k-1}(1/j^2))}{\Sigma_{j=1}^{k-1}(1/j)} > 2\left(1 - \frac{1}{k}\right).$$

The term $2(1 - 1/k)$ is the time until a MRCA (unconditional to a recombination event). The greater the time until a MRCA, the higher the chance of a recombination nearby (Wiuf and Hein 1999). Therefore,

$$E[r_k|B_k(0) = 1] \le E[r_k].$$

For example, $E[r_2|B_2(0) = 1] = \frac{1}{2}$, but $E[r_2] = \infty$ and $E[r_3|B_3(0) = 1] = \frac{1}{3}$, but $E[r_3] = \log(2) \approx 0.69$.

3. Number of MRCAs: The number of different MRCAs is upward bounded by $S_k(\rho)$, the number of segments carrying ancestral material in the set of MRCAs and hence bounded in expectation by

$$1 + \rho\left(1 - \frac{2}{k(k+1)}\right)$$

[according to (5)]. From this we find that the expectation of $S_k(\rho) - 1$ is linear in ρ. In contrast, the number of MRCAs is not likely to be linear in ρ, because each MRCA might have the ancestral material located on several segments, and these can be spread over all the sequence length. There is a chance that two positions, q_1 and q_2, when very far apart share MRCA, but are located on different segments of ancestral material (Figure 3). In Figure 4 the ratio of the expectation of MRCAs $- 1$ to the expectation of $S_k(\rho) - 1$ is plotted for increasing sequence length and different values of k. For fixed ρ the ratio quickly becomes independent of sample size. This indicates that the number of segments into which each MRCA is partitioned is almost independent of k.

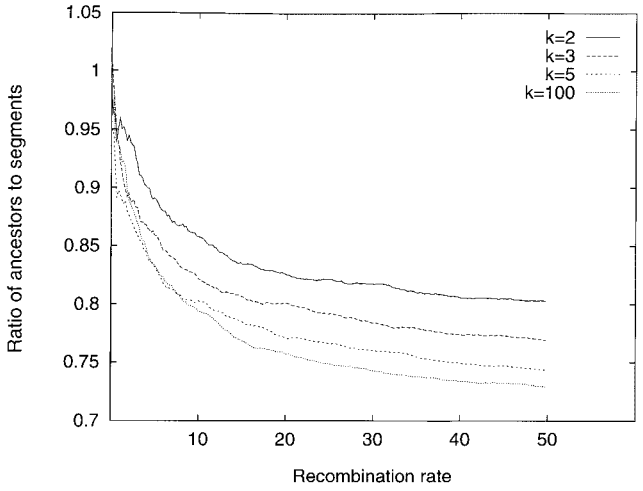

Figure 4.—The ratio of the expected number of MRCAs to the expected number of segments, $S_k(\rho)$, carrying ancestral material in the set MRCAs. One is subtracted from both numbers before taking the ratio (see Results, **3**). The ratio is a slowly decreasing function of ρ for fixed sample size, k. The small slopes of the curves for larger values of ρ indicate that the total span of positions sharing a MRCA is very narrow: Increasing sequence length will not increase the amount of positions sharing the MRCA. The curve for sample sizes 10 and higher is almost identical to the curve for $k = 100$.

4. The leftmost MRCA: Consider now the MRCA to position $q = 0$, MRCA$_k(0)$. In the case $k = 2$ we can calculate the expected amount of positions sharing MRCA with position 0, $L_2(\rho)$ [see (3)], as a function of $R \equiv \rho/2$, the sequence length. We find (see appendix)

$$E[L_2(\rho)] \approx \tfrac{1}{4}\log(18 + 26\rho + \rho^2) + 0.607.$$

The exact expression can be found in the appendix. For large ρ's, $L_2(\rho)$ increases like $\log(\rho)/2$.

Further, we find the following lower bound on the variance of $L_2(\rho)$ (see appendix):

$$\begin{aligned}\mathrm{Var}[L_2(\rho)] &= E[L_2(\rho)^2] - E[L_2(\rho)]^2 \\ &\ge \tfrac{1}{2}\rho - \tfrac{1}{2}\log(1+\rho) - E[L_2(\rho)]^2.\end{aligned}$$

Because $E[L_2(\rho)]$ is of order $\frac{1}{2}\log(1 + \rho)$, the variance is of order $\frac{1}{2}\rho$ at least.

Combining the expression for the expectation of $L_2(\rho)$ with the lower bound on the variance, we find that the normalized variable $2L_2(\rho)/\log(1 + \rho)$ has expected

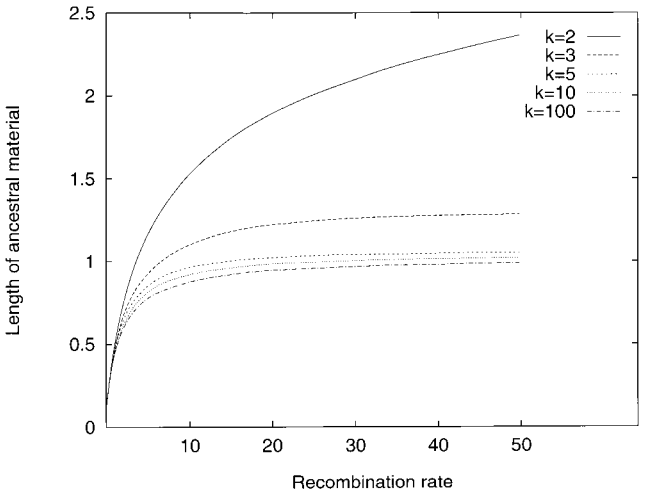

Figure 5.—Expected length of material sharing MRCA with position 0 in the sample. The MRCA to position 0 is called the leftmost MRCA. The number k denotes sample size. For $k = 2$, the curve grows like $\log(1 + \rho)/2$, whereas for larger k values, the curve becomes almost constant. This indicates that increasing the sequence length will not increase the amount of positions sharing MRCA with position 0. This observation is in concordance with Figure 4.

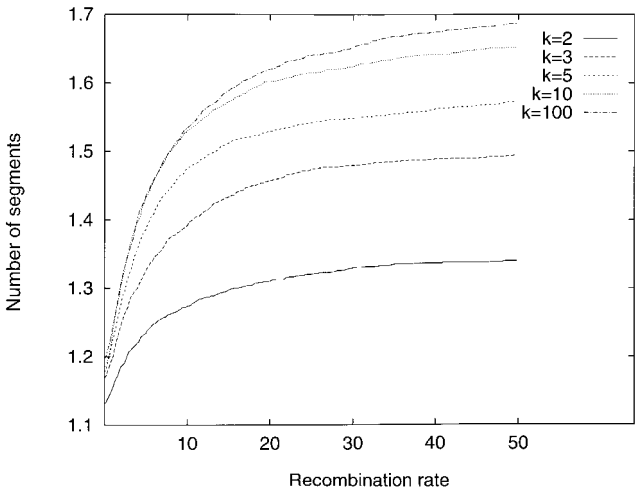

Figure 6.—Expected number of segments on leftmost MRCA. Each segment consists of positions sharing MRCA with position 0 in the sample. The material in between two segments is either nonancestral to the sample or ancestral. In the latter case, the MRCA of a position q is different from the MRCA to position 0. The number of segments increases as function of k, but the total length of the segments decreases in k (see Figure 5).

value $\sim$1, but has a variance that increases without bound for increasing sequence length. Thus, there is very large variation in the amount of positions sharing MRCA with position 0 when sample size is 2.

For larger sample sizes, $k > 2$, we have $P\{\text{MRCA}_k(0) = \text{MRCA}_k(q)\} \leq P\{\text{MRCA}_2(0) = \text{MRCA}_2(q)\}$ (see appendix), and hence that

$$E[L_k(\rho)] \leq E[L_2(\rho)] \approx \tfrac{1}{2}\log(1 + \rho).$$

This bound is very crude. We have simulated the length, $L_k(\rho)$, for samples of different sizes and found that for large sample sizes, the expected length is a slowly growing function of ρ (Figure 5), and for $k > 2$ it almost becomes constant. The difference in expected length between samples of size 10 and 100 is $<5\%$ within the range ρ is varied. This indicates a quick convergence of expected length for increasing sample sizes and fixed ρ.

The set of positions sharing MRCA with position 0 is (potentially) partitioned into several segments (Figure 3). Figure 6 shows the expected number of such segments on the leftmost MRCA. For small sample sizes there are less segments than for large sample sizes. Moreover, as shown in the previous figure, the expected length of positions sharing MRCA with position 0, $E[L_k(\rho)]$, is smaller for large sample sizes than for small sample sizes. This supports the conclusion that the material sharing MRCA with position 0 is chopped into more segments for large sample sizes than for small sample sizes, and that these segments tend to be shorter for large samples than for smaller samples.

The histograms in Figures 7 and 8 show the amount of trapped material on the leftmost MRCA. Trapped material is between the segments sharing MRCA with position 0 (Figure 3). Figure 3 supports the conclusion that as sample size increases so does the chance of finding trapped material on the leftmost MRCA, *i.e.*, the chance that the ancestral material is located on different segments increases. We conclude, as we did in Figure 6, that the chance of several segments being on the leftmost MRCA is highest for large samples.

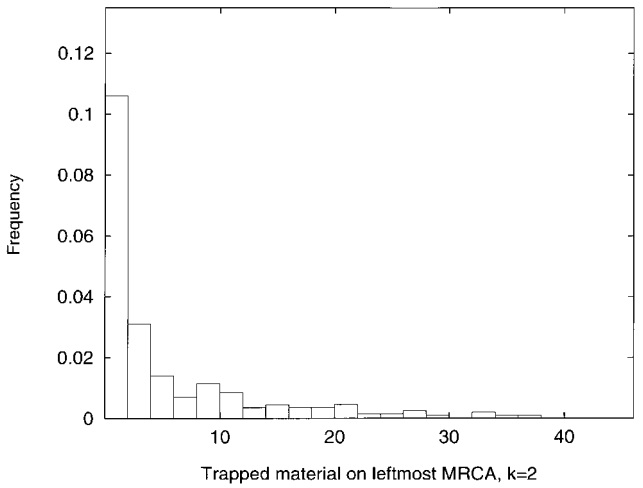

Figure 7.—Trapped material on the MRCA to position 0 in the sample (leftmost MRCA). Sample size is 2. The figure is similar to Figure 8. The frequency of leftmost MRCAs with a given value of trapped material (x-axis) is shown. The number of leftmost MRCAs without any trapped material is $\approx$0.79 (not shown), by far the most common situation. Large amount of trapped material means that the leftmost MRCA consists of several segments separated by large distances.

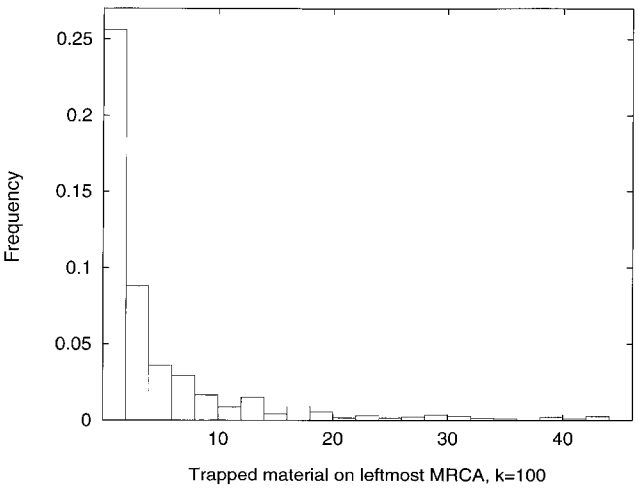

Figure 8.—Trapped material on the MRCA to position 0 in the sample (leftmost MRCA). Sample size is 100. The figure is similar to Figure 7. The frequency of leftmost MRCAs with a given value of trapped material (x-axis) is shown. The number of leftmost MRCAs without any trapped material is ≈0.51 (not shown). As the sample size goes up, the chance gets bigger that the material sharing MRCA with position 0 is spread on several segments: The number of leftmost MRCAs without any trapped material is ≈0.79 for sample size 2.

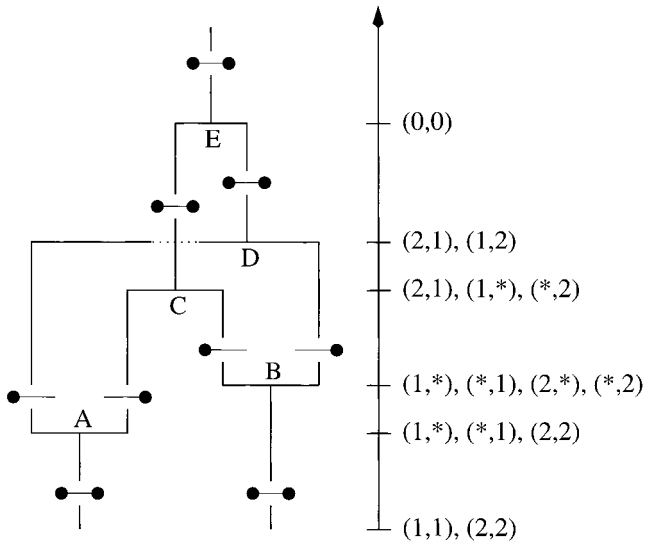

Figure 9.—Shared sequence ancestry. The figure (identical to Figure 1) shows the genealogical history of two positions. The arrow line indicates time, and the configurations of the ancestral samples are shown immediately after an event. After event D the two pairs of positions are located on two sequences, but the positions do not share sequence ancestry at this point: Both positions on both sequences are ancestral to the sample, but the positions have been swapped. The total time the positions share sequence ancestry is from time 0 until event A happens.

5. Shared sequence ancestry: Consider two positions. As the distance between the positions gets larger, their genealogical histories become less correlated. In general, the correlation between genealogies might be measured in different ways according to what aspects of the genealogy are of interest. Kaplan and Hudson (1985) found that the covariance between the total branch lengths, $G_k(q_1)$ and $G_k(q_2)$, of the genealogies in positions q_1 and q_2 is about $k/(4(k-1)R)$, where $R \equiv \rho/2$ is the distance between q_1 and q_2. A similar result will hold for the tree heights of the genealogies.

If the positions are completely linked, the positions ancestral to q_1 and q_2 are on the same ancestral sequence. When recombination is present, the ancestral positions to q_1 and q_2 will not necessarily share sequence, but might be on different sequences. The time they share ancestral sequences is a measure of the correlation between the two genealogies. We define and discuss a notion of shared sequence ancestry in this context.

Let a sample of size 2 be given. Fix two positions, q_1 and q_2, on the sequences s_1 and s_2 with distance R (recombination rate $\rho = 2R$).

Denote an ancestral state to the sample by a list $((x_{i1}, x_{i2})|i = 1, \ldots, n)$, where n is the number of ancestral sequences, and (x_{i1}, x_{i2}) denotes an ancestral sequence. The variable x_{ij} is * if position q_j on the sequence represented by (x_{i1}, x_{i2}) is nonancestral to any position in the sample, 0 if ancestral to q_j on both s_1 and s_2, 1 if ancestral to q_j on s_1 only, or 2 if ancestral to q_j on s_2 only. The definition is illustrated in Figure 9. A present-day sample is represented by ((1,1), (2,2)). We say that the positions share sequence ancestry whenever the ancestral state is $((x_{i1}, x_{i2})|i = 1, \ldots, n) = ((1,1), (2,2))$ or $= ((0,0))$. This implies that the positions ancestral to q_1 and q_2 on s_1 share an ancestor, and at the same time the ancestral positions to q_1 and q_2 on s_2 share an ancestor, possibly the same.

If the positions are completely linked ($R = 0$) the ancestral state is ((1,1), (2,2)) until the positions find a MRCA and the state becomes ((0,0)). If the positions are less linked, the ancestral state might be different from ((1,1), (2,2)). As a measure of shared sequence ancestry we take the expectation of the time T_S spent in the state $S = ((1,1), (2,2))$ compared to the expectation of the time T_j, $j = 1, 2$ until a position finds a MRCA, *i.e.*,

$$\frac{E[T_S]}{E[T_j]} = E[T_S],$$

because $E[T_1] = E[T_2] = 1$. Standard Markov chain analysis (see appendix) gives

$$E[V_S] = \frac{1+\rho}{18+13\rho+\rho^2}\left(18+\rho-\frac{2\rho^2}{(3+\rho)(2+\rho)}\right),$$

where V_S is the number of times state S is visited (the initial time included), and hence

$$E[T_S] = E[E[T_S|V_S]] = \frac{1}{1+\rho}E[V_S]$$
$$= \frac{1}{18+13\rho+\rho^2}\left(18+\rho-\frac{2\rho^2}{(3+\rho)(2+\rho)}\right),$$

because $T_S|V_S \sim \Gamma(V_S, 1+\rho)$. If recombination is not present so that q_1 and q_2 are completely linked, then $E[T_S]/E[T_j] = 1$. As ρ increases, $E[T_S]/E[T_j]$ decreases toward 0. The genealogies of the two positions become less correlated as the chance of a recombination break between the two positions increases; for very high recombination rate $\rho = 2R$ the two positions ancestral to q_1 and q_2 on s_1 are on the same sequence with probability $\approx 1/(1+R)$ (Wiuf and Hein 1997). Note that this measure of correlation between the genealogies in two distinct positions is of order $1/\rho$.

Similarly, we can calculate the shared sequence ancestry given that the two positions find a MRCA at the same time, *i.e.*, given $\text{MRCA}_2(q_1) = \text{MRCA}_2(q_2)$ or $T_1 = T_2$. We find

$$E[T_j|T_1 = T_2] = \frac{3(6+\rho)}{18+13\rho+\rho^2}$$

and

$$\frac{E[T_S|T_1 = T_2]}{E[T_j|T_1 = T_2]} = \frac{1}{3(6+\rho)}\left(18+\rho-\frac{2\rho^2}{(3+\rho)(2+\rho)}\right)$$

by similar analysis to that above (see appendix). Whereas $E[T_S]/E[T_j]$ decreases from 1 toward 0 for ρ increasing, $E[T_S|T_1 = T_2]/E[T_j|T_1 = T_2]$ decreases from 1 toward $\frac{1}{3}$.

The value of $E[T_j|T_1 = T_2]$ is $\sim 3/\rho$ for large ρ's. The time spent in S is only about $\frac{1}{3}$ the total time, so that in general several events happen before the MRCA and not just a single coalescent event.

6. Tree heights and branch lengths: Let $T_k(q)$ denote the time until a MRCA in position q in a sample of size k, and $G_k(q)$ the total branch length of the genealogy in position q. Because one position cannot be subject to recombination these two variables depend on only the coalescent process and not the recombination process. The expectation of $T_k(q)$ is

$$E[T_k(q)] = 2\left(1-\frac{1}{k}\right) < 2.$$

However, the distribution of $T_k = \max_{0\le q\le \rho/2} T_k(q)$ is highly dependent on the recombination rate ρ. The variable T_k is the time until all positions along the sequences have found a MRCA. Griffiths and Marjoram (1997) found that the expectation of T_k is bounded:

$$0 \le E[T_k] - E[T_k(0)] \le \frac{\rho}{2}.$$

This bound is uniform in k, and linear in ρ. We have simulated T_k to see how good this bound is (Figure 10). $E[T_k] - E[T_k(0)]$ seems to converge toward a logarithmic limit.

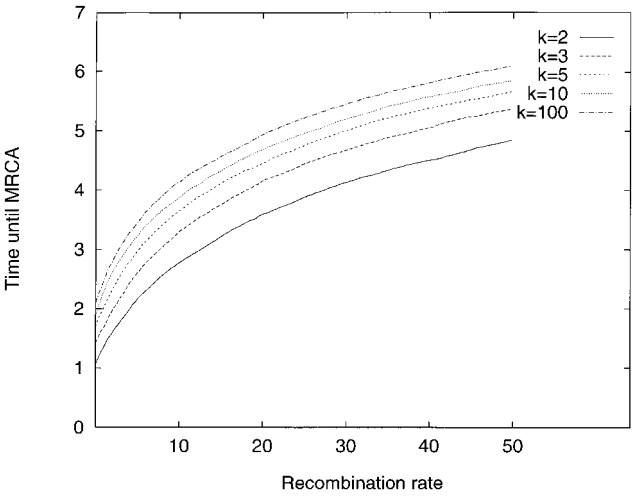

Figure 10.—The expected time until all positions have found a MRCA. This expected time becomes quickly independent of sample size: The difference between sample size 25 and sample size 100 is <2%. For $\rho = 0$, the time until a MRCA is distributed according to the coalescent process and the expectation is $2(1-1/k)$, k denoting sample size.

Similarly, the expectation of $G_k(q)$ is

$$E[G_k(q)] = 2\sum_{i=1}^{k-1}\frac{1}{i}.$$

Using the technique of Griffiths and Marjoram (1997, Theorem 3.1), it can be proved (see appendix) that

$$0 \le E[G_k] - E[G_k(0)] = E[\max_{0\le q\le p/2} G_k(q)]$$
$$-E[G_k(0)] \le 4\rho.$$

The variable G_k is the maximum of all total branch lengths. On the basis of simulations presented in Figure 11 it is obvious that $E[G_k] - E[G_k(q)]$ does not depend linearly on ρ, but seems to converge toward a logarithmic limit.

DISCUSSION

We have discussed properties of the ancestry of k sequences sampled from the coalescent process with recombination. We have done so mainly by simulations of sample histories. A number of variables derived from the genealogies were observed, and the expectations over 2000 simulations were calculated.

Each of these variables describes an aspect of the ancestry of a sample. One should in general be cautious in interpreting the behavior of a process from expectations only: The variation in the process is ignored when relying on expected values, and it is not guaranteed that

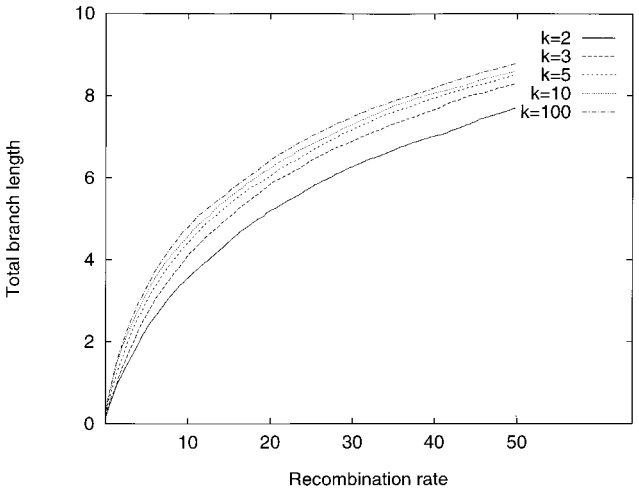

Figure 11.—The expectation of the maximum of the total branch length over all positions in the sample. We have subtracted the expectation of the total branch length in a single position, *i.e.*, subtracted $2\sum_{i=1}^{k-1}(1/i)$, where k denotes sample size. This expected time becomes quickly independent of sample size: The difference between sample size 10 and sample size 100 is <3%. For $\rho = 0$, the expectation is zero.

the expected value represents a "typical outcome" of the variable.

However, a comparison of expected values for varying sample sizes, k, and varying recombination rates, ρ, gives an idea of how the variables depend on k and ρ and thereby an idea of the amount of information in the size of the sample.

In particular we were interested in the leftmost MRCA, *i.e.*, the MRCA to position 0 in the sample. The length of material sharing MRCA with position 0 decreases with increasing sample size toward a limit (sample size ∞). At the same time, the number of segments on the leftmost MRCA increases with increasing sample sizes.

For samples of size 2, we discussed a concept of sharing sequence ancestry between two positions (or loci). It was shown that, even when the two positions share a MRCA, the proportion of time they share ancestral sequences is short. For an increasing recombination rate the positions share sequence ancestry in $\frac{1}{3}$ of the time until a MRCA.

Finally, simulations indicated that the expectations of the variables $T_k - T_k(q)$ and $G_k - G_k(q)$ are bounded in k (sample size) by a logarithmic function of ρ. The variable T_k is the time until all positions in the sample have found a MRCA, and G_k is the maximum of the total branch length of the genealogy over all positions. The bounds revived theoretically are far higher than the simulated curves.

It is interesting that the structure of a MRCA to a sample of sequences converges very quickly toward a limit structure (in expectations). In all figures, the difference between the simulation results for samples of size 10 (in a few cases 25) and larger sample sizes is close to zero. This indicates that the structure of a MRCA of a sample of size 10 has identical structure to a MRCA of the whole population.

Moreover, the expectation of the waiting time until all positions in a sample of size 10 have found a MRCA is about the same as the expectation of the waiting time until all positions in the whole population have found a MRCA. This finding is very similar to a result about the coalescent without recombination: The distribution of the waiting time until a sample of size 10 has found a MRCA is almost distributed like the same waiting time until the whole population has found a MRCA.

The explanation for this seems to be the following: Consider a large sample. The time during which there are many ancestors to the sample is considerably smaller than the time during which there are a few ancestors only. The rate of recombination is $k\rho/2$ if there are k sequences, and the rate of coalescence is $k(k-1)/2$. If k is much larger than ρ, most events will in the beginning be coalescence events. Thus, the time from the present until the whole sample has been reduced to a small number of ancestors by coalescence events will be distributed similarly to the time until a large sample is reduced to a small sample in the coalescent without recombination. It is, therefore, the size of the minor number of ancestors that determines the structure of the variables we have discussed.

However, it is surprising that the convergence in sample size k seems almost uniform in ρ. The reason for this might be that the range within which ρ has been varied is too narrow to detect the dependence on ρ.

As an example, consider a large sample of human DNA sequences. Assume that the probability r of a recombination event between two nucleotides per generation per sequence is 10^{-7} and that the effective population size of the human population is $2N = 10^4$. If the number of nucleotides is $L = 10^4$ (typical gene length), then $\rho = 4NLr = 2 \cdot 10^4 \cdot 10^4 \cdot 10^{-7} = 20$ and sequence length is $R = \rho/2 = 10$. In this case, there are $\sim$15 different MRCA consisting of $\sim$21 segments in total [see (5) and Figure 4]. Each segment will on average be $E[s_k|A_k(0) = 1] \cdot L/R = \frac{1}{2} \cdot 10^4/10 = 500$ [see (7)] nucleotides long and each MRCA 700 nucleotides long ($L/15 = 10^4/15$). Focus now on nucleotide 500 in the sequences. The length of the sequences to the right of the nucleotide is $R/2 = 5$. From Figure 5 we find that the expected number of positions sharing MRCA with nucleotide 500 is about $0.75 \cdot 10^4/10 = 750$. Similarly, $\sim$750 nucleotides to the left of number 500 will share MRCA with nucleotide 500; in total, 1500 nucleotides.

The expected time back until all nucleotides have found a MRCA is $\sim 5 \cdot 2N = 50{,}000$ generations (Figure 10). Counting 1 generation as 20 years, this is about 1 million years ago, whereas a random spot has average time to the MRCA of 40,000 years.

We thank Mikkel Nygaard Hansen for help with implementation

of the simulation program. Bernt Guldbrandtsen is thanked for reading and commenting on the manuscript. J.H. was supported by Danish Research Council grant SNF 94-0163-1.

LITERATURE CITED

Daley, D. J., and D. Vere-Jones, 1988 *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.

Griffiths, R. C., 1991 The two-locus ancestral graph, in *Selected Proceedings of the Symposium of Applied Probability, Sheffield 1989, IMS Lecture Notes—Monograph Series, 18*, edited by I. V. Basawa and R. L. Taylor, Hayward, CA.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comp. Biol. **3/4:** 479–502.

Griffiths, R. C., and P. Marjoram, 1997 An ancestral recombination graph, pp. 257–270 in *Progress in Population Genetics and Human Evolution, IMA Volumes in Mathematics and its Applications*, Vol. 87, edited by P. Donnelly and S. Tavaré. Springer-Verlag, Berlin.

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

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

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

Kemeny, J. G., and J. L. Snell, 1960 *Finite Markov Chaines.* Van Nostrand Company, New York.

Kingman, J. F. C., 1982 The coalescent. Stoch. Process. Appl. **13:** 235–248.

Simonsen, K. L., and G. A. Churchill, 1997 A Markov chain model of coalescence with recombination. Theor. Popul. Biol. **52:** 43–59.

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

Wiuf, C., and J. Hein, 1997 On the number of ancestors to a DNA sequence. Genetics **147:** 1459–1468.

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. (in press).

Communicating editor: R. R. Hudson

APPENDIX

Numbers refer to sections in results.

2. Following Daley and Vere-Jones (1988), the expressions $E[r_k|B_k(0) = 1]$ and $E[s_k|A_k(0) = 1]$ are understood in the following sense. Extend the sequences by a small interval I of length δ (in units of expected number of recombinations per sequence per $2N$ generations) to the left of the position 0. Let $\beta_k(\delta)$ denote the event in which there is a least one recombination event within the sequence interval I in the history of the sample of size k. Similarly, let $\alpha_k(\delta)$ denote the event in which there is at least one shift in MRCAs within the interval I. Then

$$E[r_k|B_k(0) = 1] = \lim_{\delta\to 0} E[r_k|\beta_k(\delta)]$$

and

$$E[s_k|A_k(0) = 1] = \lim_{\delta\to 0} E[s_k|\alpha_k(\delta)].$$

According to Daley and Vere-Jones (1988), the conditional expectations are given by

$$E[r_k|B_k(0) = 1] = \frac{\rho}{2E[R_k(\rho)]}$$

and

$$E[s_k|A_k(0) = 1] = \frac{\rho}{2(E[S_k(\rho)] - 1)},$$

if the processes are stationary and the numbers $R_k(\rho)$ and $S_k(\rho)$ become infinite with probability one for sequences of infinite length, *i.e.*, $R_k(\rho)$, $S_k(\rho) \to \infty$ as $\rho \to \infty$. The denominators are the expectations of $R_k(\rho)$, $S_k(\rho) - 1$, respectively, and the numerators are the sequence length $R = \rho/2$. Assume sample size k is finite. Because $S_k(\rho) - 1 \le R_k(\rho)$, it suffices to consider $S_k(\rho)$ only,

$$\{S_k(\rho) < \infty\} = \bigcup_{i=n}^{\infty}\{S_k(\rho) < i\} \subseteq \bigcup_{q=p}^{\infty}\{T_k(q) = T_k(q + q^2)\}$$

for arbitrary n and p natural numbers, and where $T_k(p)$ denotes the height of the local tree in position p. The probability of $\{T_k(q) = T_k(q + q^2)\}$ is bounded by $(k - 1)k/2\, P(q^2)$, where $P(r)$ is the probability that two positions on two sequences separated by a distance r find a MRCA at the same time. Griffiths (1991) found

$$P(r) = \frac{9 + r}{9 + 13r + 2r^2} \le \frac{2}{3r},$$

and therefore

$$P(S_k(\rho) < \infty) \le \sum_{q=p}^{\infty} P\{T_k(q) = T_k(q + q^2)\}$$

$$\le \sum_{q=p}^{\infty} \frac{k(k-1)}{3q^2} \to 0$$

for $p \to \infty$. Hence $S_k(\rho)$ and $R_k(\rho)$ become infinite with probability one and the conditional expectations are given by (6) and (7) for $k < \infty$.

Consider now the case $k = \infty$. The chance that there is at least one recombination event in the history of the sample in any interval of the sequences is one, independent of the size of the interval considered. That is, $P(\beta_k(\delta)) = 1$ for all $\delta > 0$ and $P(R_\infty(\eta) \ge 1) = 1$ for all $\eta > 0$. Thus,

$$E[r_\infty|\beta_k(\delta)] = E[r_\infty] = E[r_\infty|R_\infty(\eta) \ge 1] \le \eta.$$

As this holds for all $\eta > 0$, we conclude that

$$E[r_\infty|B_\infty(0) = 1] = 0$$

as desired.

To prove $E[s_\infty|A_\infty(0) = 1] = \frac{1}{2}$ we show that $E[S_\infty \cdot (\rho)] = 1 + \rho$. Unfortunately, $S_k(\rho)$ does not converge toward $S_\infty(\rho)$ in any regular way. Label all sequences in an infinite sample by numbers. Let $S_k^*(\rho)$ be the number of recombination breaks, q, in the history of the first k sequences such that $\text{MRCA}_\infty(q - \varepsilon) \neq \text{MRCA}_\infty(q + \varepsilon)$ provided ε is small. Clearly

$$1 + S_k^*(\rho) \le S_k(\rho) \quad \text{and} \quad 1 + S_k^*(\rho) \uparrow S_\infty(\rho).$$

Because $1 + E[S_k^*(\rho)] \le E[S_k(\rho)] \le 1 + \rho$, then $E[S_\infty(\rho)] \le 1 + \rho$, and $S_\infty(\rho)$ is finite almost surely for all ρ. This implies that

$$S_\infty(\rho) = 1 + \lim_{\varepsilon \to 0} \sum_{i=1}^{R\varepsilon^{-1}} \cdot 1\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\}.$$

Further, $1\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\} \le 1\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((2i+1)\varepsilon/2)\} + 1\{\mathrm{MRCA}_\infty \cdot ((2i+1)\varepsilon/2) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\}$, so by dominated convergence

$$E[S_\infty(\rho)] = 1 + \lim_{\varepsilon \to 0} \sum_{i=1}^{R\varepsilon^{-1}} \cdot P\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\}.$$

Consider the positions $i\varepsilon$ and $(i+1)\varepsilon$ for fixed ε. We have

$$1\{\mathrm{MRCA}_k(i\varepsilon) \neq \mathrm{MRCA}_k((i+1)\varepsilon)\} \to 1\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\}$$

for $k \to \infty$ and all $\varepsilon \ge 0$, and therefore

$$P\{\mathrm{MRCA}_k(i\varepsilon) \neq \mathrm{MRCA}_k((i+1)\varepsilon) = 2\varepsilon\left(1 - \frac{2}{k(k+1)}\right) \to P\{\mathrm{MRCA}_\infty(i\varepsilon) \neq \mathrm{MRCA}_\infty((i+1)\varepsilon)\}.$$

The equality is given by Griffiths and Marjoram (1997). It follows that

$$E[S_\infty(\rho)] = 1 + \lim_{\varepsilon \to 0} \sum_{i=1}^{R\varepsilon^{-1}} 2\varepsilon = 1 + \rho$$

and hence

$$E[s_\infty | A_\infty(0) = 1] = \frac{\rho}{2\rho} = \frac{1}{2}$$

as desired.

4. By definition of $L_2(\rho)$ we have $L_2(\rho) \le \rho/2$. Thus, by dominated convergence

$$\mathbf{E}[L_2(\rho)] = E[\lim_{\varepsilon \to 0} \varepsilon \sum_{i=1}^{R\varepsilon^{-1}} 1\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(i\varepsilon)\}] = \lim_{\varepsilon \to 0} \varepsilon \sum_{i=1}^{R\varepsilon^{-1}} P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(i\varepsilon)\}.$$

We have $\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(q)$ if $T_2(0) = T_2(q)$, where $T_k(q)$ denotes the time until a MRCA in position q. Using the expression in Griffiths (1991, Equation 2.12), for the probability of $T_2(0) = T_2(q)$, we conclude that

$$E[L_2(\rho)] = \int_0^{\rho/2} \frac{9 + x}{9 + 13x + 2x^2} dx = \frac{1}{4}\log(18 + 26\rho + \rho^2) - \frac{23}{2\sqrt{97}} \coth^{-1}\left(\frac{13 + 2\rho}{\sqrt{97}}\right) - \frac{1}{4}\log(18) + \frac{23}{2\sqrt{97}}\coth^{-1}\left(\frac{13}{\sqrt{97}}\right) \approx \frac{1}{4}\log(18 + 26\rho + \rho^2) + 0.607.$$

Regarding the variance of $L_2(\rho)$, we have

$$L_2(\rho)^2 = 2 \lim_{\varepsilon \to 0} \varepsilon^2 \sum_{i<j} \cdot 1\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(i\varepsilon) = \mathrm{MRCA}_2(j\varepsilon)\}.$$

Because $L_2(\rho)^2 \le \rho^2/4$, the variance is given by

$$\mathrm{Var}[L_2(\rho)] = E[L_2(\rho)^2] - E[L_2(\rho)]^2 = 2 \lim_{\varepsilon \to 0} \varepsilon^2 \sum_{i<j} \cdot P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(i\varepsilon) = \mathrm{MRCA}_2(j\varepsilon)\} - E[L_2(\rho)]^2.$$

The probability $P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(i\varepsilon) = \mathrm{MRCA}_2(j\varepsilon)\}$ is bounded from below by

$$\frac{1}{1 + 2i\varepsilon + 2(j-i)\varepsilon} = \frac{1}{1 + 2j\varepsilon}.$$

This is the probability that the first event going backward in time is a coalescent event, whereby the three positions find a MRCA at the same time. Hence,

$$\mathrm{Var}[L_2(\rho)] \ge 2 \int_0^R dx \int_0^{R-x} \frac{1}{1 + 2x + 2y} dy = \frac{1}{2}\rho - \frac{1}{2}\log(1 + \rho) - E[L_2(\rho)]^2.$$

To prove $P\{\mathrm{MRCA}_k(0) = \mathrm{MRCA}_k(q)\} \le P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(q)\}$ we proceed as follows: If $\mathrm{MRCA}_k(0) = \mathrm{MRCA}_k(q)$, then the last event must be a coalescence between two sequences, both carrying ancestral positions to 0 and q. Let S denote the state consisting of two ancestral sequences and both positions 0 and q ancestral to the sample. Let s denote the time until state S is entered for the first time, and let $F(\cdot)$ denote the probability distribution of s. We have

$$\begin{aligned} P\{\mathrm{MRCA}_k(0) = \mathrm{MRCA}_k(q)\} &= \int_0^\infty P\{\mathrm{MRCA}_k(0) = \mathrm{MRCA}_k(q) | s\} dF(s) \\ &= P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(q)\} \int_0^\infty dF(s) \\ &= P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(q)\} P(s < \infty) \\ &< P\{\mathrm{MRCA}_2(0) = \mathrm{MRCA}_2(q)\} \end{aligned}$$

as desired.

5. The state space of the Markov chain consists of all

possible ancestral configurations $((x_{i1}, x_{i2})|i = 1, \ldots, n)$ to the sample. The variables x_{ij} are defined in **5.** The state ((0,0)) is absorbing. This is very similar to a Markov chain in Simonsen and Churchill (1997) describing a two-locus model with sample size 2. The difference is that we distinguish between some states, *e.g.*, ((1,1), (2,2)) and ((1,2), (2,1)), while Simonsen and Churchill (1997) do not. The transition probabilities are given by expressions similar to expressions in Simonsen and Churhill (1997), and follow from the structure of the two-locus model. For example, the probability of going from state ((1,2), (1,2)) to state ((1,*), (*,2), (2,1)) is

$$P\{((1,*),(*,2),(2,1))|((1,2),(1,2))\} = \frac{1}{1+\rho}.$$

This happens only if the first event is a recombination event. The expectation of the number of times state $S =$ ((1,1), (2,2)) is visited can then easily be found using standard Markov chain techniques [see, *e.g.*, Kemeny and Snell (1960), theorem 3.5.4]. This gives the desired result.

To obtain the expectations $E[T_s|T_1 = T_2]$ and $E[T_j|T_1 = T_2]$, first note that for states S' and S'' we have

$$P(S''|S', T_1 = T_2) = P(S''|S') \frac{P(T_1 = T_2|S'')}{P(T_1 = T_2|S')}$$

and that the conditional chain is Markov. The probability $P(S''|S')$ is known from the unconditional chain, and $P(T_1 = T_2|S')$ and $P(T_1 = T_2|S'')$ can be found using, *e.g.*, Kemeny and Snell (1960), theorem 3.3.7. Applying Kemeny and Snell (1960), theorem 3.5.4, to the conditional Markov chain gives us the expectations of number of times $V_{S'}$ a state S' is visited. This is similar to the calculations above in the unconditional case. The expressions of the conditional expectations are now consequences of the following: (1) The event $\{T_1 = T_2\}$ depends on the jump chain of the process only, and not on the time between jumps and (2) $T_{S'}$ conditional on $V_{S'}$ is $\Gamma(V_{S'}, \lambda_{S'})$ distributed. The parameter $\lambda_{S'}$ is the intensity parameter of the exponential waiting time until the chain leaves state S'.

6. Similar to Griffiths and Marjoram (1997, Theorem 3.1), we find

$$E[G_k] - E[G_k(0)] \le E[R_k(\rho)]\, E[G_l - G_r|G_l > G_r] \cdot P(G_l > G_r), \tag{10}$$

where $P(G_l > G_r)$ is the probability that the total length of the genealogy G_l just to the left of a recombination breakpoint p is larger than the length of the genealogy G_r just to the right of p. Assume a recombination event happens while there are j ancestors, $j = 2, \ldots, k$ to the sample. Just after the recombination event there will be $j + 1$ ancestors to the sample including the two recombined sequences, s_l and s_r. The sequence s_l is ancestral to the positions just to the left of p, and s_r is ancestral to the positions just to the right of p. The probability $p(j, i, h)$ that s_l coalesces to a lineage other than the lineage of s_r while there are i ancestors, $i = 3, \ldots, j + 1$, and that s_r coalesces while there are h ancestors, $h = 2, \ldots, i - 1$, is given by

$$p(j,i,h) = \frac{4(i-2)(h-1)}{(j+1)j^2(j-1)}.$$

By conditioning further on (j, i, h) in (10) and noting that $E[G_l - G_r|G_l > G_r, j, i, h] < 4/(h - 1)$, we find that

$$E[G_k] - E[G_k(0)] \le E[R_k(\rho)] \sum_{j,i,h} \frac{4(i-2)(h-1)}{(j+1)j^2(j-1)} \cdot \frac{4}{h-1} \frac{1}{(j-1)\sum_{l=1}^{k-1} 1/l},$$

where the last term is the probability that the recombination, given that it occurs, happens while there are j ancestors (Griffiths and Marjoram 1997). Reducing the sum, we find

$$E[G_k] - E[G_k(0)] \le E[R_k(\rho)] \frac{4}{\sum_{l=1}^{k-1} 1/l} = 4\rho$$

as desired.