DOI: 10.1534/genetics.107.071126

On Recombination-Induced Multiple and Simultaneous Coalescent Events

Joanna L. Davies,[1] František Simančík, Rune Lyngsø, Thomas Mailund and Jotun Hein

Department of Statistics, University of Oxford, Oxford, OX1 3TG, United Kingdom

Manuscript received January 18, 2007
Accepted for publication October 2, 2007

ABSTRACT

Coalescent theory deals with the dynamics of how sampled genetic material has spread through a population from a single ancestor over many generations and is ubiquitous in contemporary molecular population genetics. Inherent in most applications is a continuous-time approximation that is derived under the assumption that sample size is small relative to the actual population size. In effect, this precludes multiple and simultaneous coalescent events that take place in the history of large samples. If sequences do not recombine, the number of sequences ancestral to a large sample is reduced sufficiently after relatively few generations such that use of the continuous-time approximation is justified. However, in tracing the history of large chromosomal segments, a large recombination rate per generation will consistently maintain a large number of ancestors. This can create a major disparity between discrete-time and continuous-time models and we analyze its importance, illustrated with model parameters typical of the human genome. The presence of gene conversion exacerbates the disparity and could seriously undermine applications of coalescent theory to complete genomes. However, we show that multiple and simultaneous coalescent events influence global quantities, such as total number of ancestors, but have negligible effect on local quantities, such as linkage disequilibrium. Reassuringly, most applications of the coalescent model with recombination (including association mapping) focus on local quantities.

KINGMAN (1982) models the ancestry of a sample of sequences with a continuous-time Markov process referred to as the Kingman coalescent. Lineages collide or coalesce after random exponential waiting times with rate dependent upon the population and sample size. This means that the probability of multiple (*i.e.*, three or more sequences coalescing into a common ancestor in a single coalescent event) and simultaneous (*i.e.*, two or more coalescent events happening at exactly the same time) coalescent events is zero. The derivation of the process can be obtained by scaling the discrete-time Wright–Fisher model and taking the limit as the population size tends to infinity. This model is extended by HUDSON (1983) to incorporate recombination. The derivation of Hudson's continuous-time approximation to the Wright–Fisher model with recombination is discussed later in more detail but is valid provided only that the set of ancestors to the sample of extant sequences remains small relative to the effective population size. In such situations it is justified to assume that multiple and simultaneous coalescent events do not occur in the evolutionary history of the sample and that ancestral sequences can recombine only with nonancestral sequences and never with each other. As the sample size increases relative to the population size, the probability of such events occurring becomes nonnegligible and consequently in these instances the rate of coalescence is underestimated by Hudson's continuous-time model. Hudson's model is widely used in population genetics to describe ancestries of sequences that can recombine. Consequently it is of interest to question to what extent the rate of coalescence is underestimated and how this influences other features of the coalescent.

FU (2006) shows the Kingman coalescent KINGMAN (1982) provides a good approximation to the discrete-time Wright–Fisher Model in most cases, even when the sample size is not small relative to the population size. This study is performed in the absence of recombination and any large sample will quickly coalesce to a small sample such that the assumption soon becomes valid and the corresponding results are accurate. In the presence of recombination this is not the case; the process tracking the number of sequences ancestral to the extant sample can be shown to reach an equilibrium distribution in which the number of sequences remains large for a significant amount of time.

PITMAN (1999), SAGITOV (1999), SCHWEINSBERG (2000), and SAGITOV (2003) derive continuous-time exact coalescent processes allowing for coalescents with multiple collisions, simultaneous multiple collisions, and simultaneous and multiple collisions, respectively, although none of these processes incorporate recombination. WIUF and HEIN (1997) derive analytical results

[1] *Corresponding author:* Department of Statistics, University of Oxford, 1 S. Parks Rd., Oxford, OX1 3TG, United Kingdom.
E-mail: davies@stats.ox.ac.uk

for the expectation and the variance of the number of ancestral segments and the expected length of a segment and approximate simulation results on the mean number of ancestors to a sample of sequences subject to recombination and coalescence. These results are derived using Hudson's approximate continuous-time model with recombination.

In this article we compare results obtained by simulations of the exact Wright–Fisher coalescent with recombination with that of Hudson's continuous-time approximation. Our simulation results can be considered in two categories: quantities that are calculated locally at a segment level and quantities calculated globally. Local quantities do not require knowledge of the entire composition of each sequence (for example, where a segment is located) whereas global quantities require knowledge of the entire sequence. We show that local quantities including average segment length, the total number of segments, linkage disequilibrium, and the total length of ancestral material are well approximated by Hudson's continuous-time model whereas global quantities including the total number of sequences carrying ancestral material and the rate of coalescence differ markedly between the models.

THE ANCESTRY OF A SAMPLE OF SEQUENCES IN THE PRESENCE OF RECOMBINATION

In the following section we describe the exact discrete-time Wright–Fisher model with recombination and Hudson's continuous-time approximation of this model. We also show where assumptions are made in the derivation of the continuous-time approximation and when they may be considered inappropriate. In our results we simulate from these models and use them to calculate Monte Carlo estimates of the expectation of various local and global quantities.

The Wright–Fisher model with recombination: The basic Wright–Fisher model not including recombination is a forward-in-time model for the evolution of a haploid population of constant size. The next generation can be simulated by selecting individuals from the current generation at random (with replacement). Each time an individual is selected it becomes the parent of a new individual that is added to the next generation. Equivalently it can be viewed backward in time to simulate the genealogy of a sample or a population; ancestors to a current sample in the previous generation are selected randomly from the population of the previous generation. When two or more individuals in the next generation share a parent, their lineages coalesce and this is referred to as a coalescent event. It can be extended to incorporate more complex structures including the addition of recombination.

The exact discrete-time (in generations) Wright–Fisher model with recombination can be described as follows: Let the population of haploid individuals be of constant size $2N$ with sequence length $L + 1$ nucleotides. Let r be the probability of a recombination between any two consecutive nucleotides in a sequence per generation. Then the distribution of recombination breakpoints on a single sequence per generation is binomial with expectation rL. Since r is assumed to be very small and L large, the binomial distribution is well approximated by a Poisson distribution with intensity R, where $R := rL$. Hence it is reasonable to approximate the discrete-sequence model with a continuous-sequence model where recombination breakpoints along a sequence can be simulated by placing them along a real interval of length R according to events of a Poisson process with constant rate 1. If R is also small then R is approximately the probability of one or more recombination events occurring along the sequence in a single generation. The parameter R is usually specified directly (rather than L and r separately), and it allows R to be well defined in the limit as $r \rightarrow 0$ and $L \rightarrow \infty$ if desired.

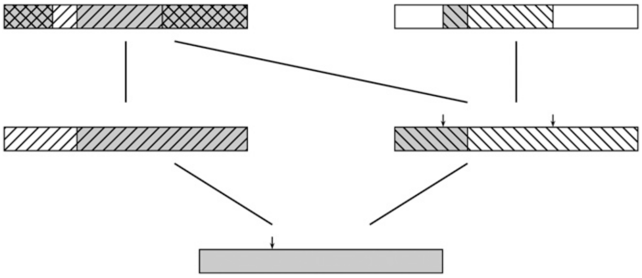

FIGURE 1.—An illustration of how ancestral material is traced two generations back in time. Material ancestral to the bottom sequence (the sample/offspring) is shaded, while material ancestral to the two parents is shown with hatched lines going in opposite directions. Although the left grandparent carries material ancestral to all of the left parent, it does not carry material ancestral to all of the offspring. Some of this has been passed via the right parent to the right grandparent. Also observe that the amount of material ancestral to the parents at the grandparents is less than twice the sequence length, as material has coalesced in the crosshatched regions of the left grandparent.

The ancestry of a sample of sequences can be simulated back in time. For each sequence we place recombination breakpoints at exponentially distributed intervals and select two parents at random from the previous generation. In Figure 1 recombination breakpoints are indicated by small arrows and the two parents selected from the previous generation are the sequences in the middle row. The genetic material of a current sequence is a mosaic of the material of its parents. The breakpoints determine from which parent the material originates, alternating at each breakpoint. The genetic material in the parents that is not passed on to the offspring is referred to as nonancestral to the offspring sequence. In Figure 1 nonancestral material in the first-generation parents is shown as boxes with single hatched lines and no shading. Genetic material at a more distant ancestor to a current sequence is defined as ancestral or nonancestral in the natural way by taking the transitive closure of these relations. In Figure 1 ancestral material in the two grandparents is identified by the shaded boxes.

When considering the ancestry of a sample of sequences we trace only events and individuals carrying material ancestral to any of the sample sequences. (In the example in Figure 1 there is a single sample sequence and material ancestral to this is shaded.) From now on, we use "ancestral material" to refer to the genetic material ancestral to a sample.

Simulating recombination events in the way described previously automatically simulates coalescent events. A multiple coalescent event occurs when more than two sequences choose the same parent. Simultaneous events occur if two or more single or multiple coalescent events occur simultaneously in a generation. Recombination events between ancestral lineages are permitted such that a sequence can be involved in a recombination event and a coalescent event at the same time.

At every stage although each individual carrying ancestral material has two parents in the previous generation each point in the sequence has exactly one parent in the previous generation. So each point from a set of extant sequences evolves according to the basic coalescent without recombination with a resulting coalescent tree that is local to that position. Consequently the genealogy relating a set of extant sequences with recombination can be considered a collection of local trees together with linkage information for ancestral segments.

Hudson's continuous-time model: Hudson (1983) derives the continuous-time coalescent model with recombination. It can be obtained by taking the limit of the discrete-time Wright–Fisher model with recombination. We outline the derivations of the distributions of the waiting times between recombination and coalescent events, highlighting the assumptions made and when it may be invalid to use them.

Recombination events and coalescent events occur independently; hence the probability that a sequence is involved in both a recombination and a coalescent event simultaneously is the product of the corresponding probabilities. In discrete time when R is small and N is large this quantity is given by (3) and $\rho := 4NR$, where ρ is twice the expected number of recombination events per generation in the population. It is derived by considering two current sequences. The probability that one experiences a recombination event and the other does not is approximately given by

$$2R(1-R) \approx 2R. \tag{1}$$

The probability that one of the recombinants coalesces with the parent of the other sequence is given by

$$1-\left(1-\frac{1}{2N}\right)^2 = \frac{1}{N}-\frac{1}{4N^2} \approx 1/N. \tag{2}$$

Hence the required product is

$$\frac{1}{N}\times 2R = \frac{1}{N}\times\frac{\rho}{2N} = \frac{\rho}{2N^2}. \tag{3}$$

For fixed ρ and for large N this quantity is negligible and consequently in the continuous-time model it is assumed that such events do not occur. This assumes that ρ is sufficiently small relative to the population size. When the rate of recombination is high and the population of fixed size this assumption may be invalid.

Using the setup above for the discrete model, we first derive the continuous-time model for recombination events. The waiting time T in generations until a recombination event occurs in a single sequence is geometric. Since the number of recombination events per sequence per generation is Poisson distributed with parameter R, the probability that a recombination event does not occur in a single generation is given by e^{-R}. Then the geometric distribution for the waiting time in generations until a recombination event follows:

$$P(T=m) = (e^{-R})^{m-1}(1-e^{-R}) \quad m \in \mathbf{N}. \tag{4}$$

To take the continuous limit, let $N\rightarrow\infty$, $R\rightarrow 0$, such that $\rho/2 = 2NR$. Hence when time is rescaled to be measured in units of $2N$ generations, the continuous waiting time T_{C} is shown to be exponentially distributed with

$$\begin{aligned} P(T_{\mathrm{C}} \leq t) &= 1-(e^{-R})^{\lfloor 2Nt\rfloor} \approx 1-(e^{-2NRt}) \\ &= 1-e^{\rho t/2}. \end{aligned} \tag{5}$$

For an individual sequence, as shown above, the continuous waiting time for a recombination event is exponentially distributed with parameter $\rho/2$ so it follows that if there are k sequences ancestral to the sample then the time until the next recombination event is exponentially distributed with parameter $k\rho/2$. Then given that a recombination event occurs, it is equally likely to occur on any of the k sequences present and the recombination breakpoint is placed uniformly along the selected sequence.

We now turn our attention to the derivation of the continuous-time model for coalescent events. The rate at which coalescent events occur in the Wright–Fisher model with recombination is the same as that of the Wright–Fisher model without recombination. A coalescent event occurs between any two sequences in a single generation with probability $1/2N$ such that the waiting time until a coalescent event occurs is geometric with mean $2N$. Since any two sequences can coalesce, it is necessary to consider the probability that with a sample of k sequences from the current population, no coalescent event occurs in a single generation. This probability is given exactly by

$$\prod_{j=1}^{k}\frac{(2N-(j-1))}{2N}=\prod_{j=1}^{k-1}\left(1-\frac{j}{2N}\right). \tag{6}$$

We can expand this product as

$$\prod_{j=1}^{k-1}\left(1-\frac{j}{2N}\right)$$
$$=1-\sum_{j=1}^{k-1}\frac{j}{2N}+O\left(\frac{1}{N^2}\right)=1-\binom{k}{2}\frac{1}{2N}+O\left(\frac{1}{N^2}\right). \tag{7}$$

Only zero and first-order terms in $1/N$ are explicitly stated; higher-order terms are gathered together in the $O(1/N^2)$ term. When the sample size is small relative to the population size $2N$ these terms contribute only negligibly to the probability and are ignored in the derivation of the Kingman coalescent. The $O(1/N^2)$ terms are indeed the sum of the probabilities of all possible multiple and simultaneous coalescent events. They are only negligible provided $k \ll 2N$. As k approaches $2N$ these terms cannot be neglected. Without even considering the exact nature of the higher-order terms, this is obvious from the fact that $1-\binom{k}{2}(1/2N)<0$ when $k>2\sqrt{N}$.

The derivation of the Kingman coalescent is based on the assumption that the waiting time while there are k ancestors to a sample is geometrically distributed with mean $2N\binom{k}{2}^{-1}$. The continuous-time coalescent process is obtained by scaling time to be measured in units of $2N$ generations and letting $2N\rightarrow\infty$. The derivation of the distribution of the continuous waiting time until a coalescent event while there are k ancestral sequences is analogous to that of the waiting time until a recombination event and yields an exponential random variable with parameter $\binom{k}{2}$. To simulate the genealogy back in time, once a coalescent time has been simulated, the pair of sequences to coalesce is chosen at random out of the possible $\binom{k}{2}$.

A recombination event with a breakpoint passing ancestral material to both parents increases the number of sequences with ancestral material by one and a coalescent event decreases the number of sequences by one. The two processes can thus be seen as competing to either increase or decrease k. An algorithm to simulate the ancestry of k sequences under the continuous model of the coalescent with recombination is described in detail in Hein *et al.* (2005).

Gene conversion: The algorithms discussed previously to simulate the ancestry of a sample of sequences place recombination breakpoints along sequences. These breakpoints result in crossover recombination events such that large segments are distributed onto two different sequences. In the human genome, it is also common to see the substitution of a small fragment of DNA from one chromosome to another. They are called homologous gene conversion events and are thought to occur more frequently than would be expected if they could occur only by drawing a very small distance between breakpoints. They can be modeled directly by adding a rate of gene conversion. Gene conversion events essentially occur independently of crossover recombination and coalescent events and they can be simulated in a similar way. In complete analogy to crossover recombinations we denote by g the probability of initiating a gene conversion between any two nucleotides and define $G := gL$ and $\gamma := 4NG$. The length of the small fragment to be transferred can be either fixed or taken from another distribution. In the human genome fragments lengths vary between 100 and 300 bases, which is approximately one millionth of the total length of the genome. We take this as the fixed length of a segment and incorporate gene conversion into the discrete Wright–Fisher model and Hudson's continuous-time model as follows.

Simulate the ancestry of a sample of n sequences under the discrete Wright–Fisher model with recombination and gene conversion:

1. Start with $k=n$ sequences each of length $R+G$.
2. For each of the k ancestral sequences in the current generation, choose two parents from the previous generation at random. If the same parent is chosen twice, no gene conversion or recombination occurs. If two distinct parents are chosen, place gene conversion and recombination events along the sequence by proceeding to step 3. Otherwise place all the ancestral material on the parent chosen twice and proceed to the next sequence.
3. Simulate intervals between breakpoints along a sequence using an exponential random variable with parameter 1. Where possible, place a breakpoint along the sequence to the right of the previous breakpoint or from the left end of the sequence if it is the first breakpoint. If the end of the interval stretches beyond the length of the sequence, go straight to step 6.
4. With probability $G/(R+G)$ it is a gene conversion event and with probability $R/(R+G)$ it is a recombination event. If it is a gene conversion event go to step 5; otherwise record the breakpoint and go back to step 3.
5. Place another breakpoint on the sequence at distance one millionth to the right of the first breakpoint where this is possible. If it is possible (*i.e.*, one millionth following the breakpoint does not stretch

beyond the length of the sequence) go back to step 3. Otherwise go to step 6.

6. Distribute the ancestral material between breakpoints alternately onto the two parents chosen at random from step 2. Update k, the current number of ancestral sequences, and proceed to the next sequence.

Simulate the ancestry of a sample of n sequences under Hudson's continuous-time model with recombination and gene conversion:

1. Start with $k = n$ sequences each of length $R + G$.
2. Simulate the time back to the next event drawing from an exponential distribution with parameter $\binom{k}{2} + k\rho/2 + k\gamma/2$.
3. Determine the type of event. With probability $(k - 1)/(k - 1 + \rho + \gamma)$ it is a coalescent event, with probability $\rho/(k - 1 + \rho + \gamma)$ it is a recombination (crossover) event and with probability $\gamma/(k - 1 + \rho + \gamma)$ it is a gene conversion event.
4. If it is a recombination or a coalescent event proceed as described in Hein *et al.* (2005, Algorithm 5). Update k, the current number of ancestors to the sample, and continue. Otherwise it is a gene conversion event. Place a breakpoint at random, uniformly along the sequence. Place another breakpoint along the sequence at a distance one millionth to the right of the initial one. Then distribute the ancestral material on two newly created ancestors. Place the ancestral material between the two breakpoints onto one of the ancestors and the remainder on the other. Update k, the current number of ancestors to the sample, and continue.

Modeling gene conversion creates more breakpoints along the sequence, therefore affecting the way in which the ancestral material is distributed on a single ancestor. Each sequence can choose only two parents, and the rate at which coalescent events occur is the same as that without gene conversion; hence we would expect the number of ancestors to a sample (in equilibrium) to remain about the same, but on each ancestor we would expect to see more (yet smaller) segments of ancestral material.

RESULTS

We investigate the effect of multiple and simultaneous coalescent/recombination events via Monte Carlo simulation. We run all simulations using a constant population size of $2N = 10{,}000$ for the exact discrete model. To investigate the effect of increasing the rate of recombination we use values of $R = 0.1$, 1, 2.5, and 36. The rate of $R = 36$ is approximately the scaled length of the human genome as estimated by Kong *et al.* (2002). To investigate the effect of increasing the sample size we run simulations for sample sizes of 500, 3000, and 8000 (all out of a population of 10,000 for the discrete model). All simulation results about the equilibrium distribution are obtained by starting with a sample size of 500 and discarding the time until virtually all positions have found a common ancestor as burn in. Subsequent approximate expectations of the quantities of interest are calculated from simulation of a further 20,000 generations and are independent of the initial sample size.

We simulate from the discrete model and the continuous approximation with and without the presence of gene conversion and report comparisons of the following: (1) the total number of sequences ancestral to a sample once these processes have reached an equilibrium, (2) the average rate at which coalescent events occur when the processes are in equilibrium, (3) the total number of ancestral segments (in equilibrium), (4) the average length of an ancestral segment (in equilibrium), (5) the rate at which the amount of ancestral material decays, and (6) the r^2 measure of correlation between the two end loci of the sequences.

Simulation results without gene conversion: *The number of ancestors to a sample:* The number of sequences carrying ancestral material at time t is a stochastic process. It is described by a Markov chain (either continuous or discrete) and these processes converge in both the discrete and the continuous case to an equilibrium distribution that is independent of the initial sample size and dependent only upon the rate of recombination. This is illustrated by Figure 2.

Each of the plots in Figure 2 corresponds to simulations run with different recombination rates (as labeled in the figure). Simulations are run with sample sizes of 500, 3000, and 8000 for both the continuous and the discrete model, which are plotted as dark gray and light gray lines, respectively. For large recombination rates the approach to the equilibrium distribution and the equilibrium distribution itself differ significantly according to the type of model used. For small recombination rates (see $R = 0.1$ in Figure 2) and relatively small sample sizes, the convergence to the equilibrium is very similar and the graphs concur. Furthermore both models converge to the same equilibrium distribution. This is not surprising since with a low recombination rate, the assumption that sequences are never involved in a coalescent and a recombination event simultaneously is reasonable and for small samples the number of ancestors to the extant sample at any point back in time does not grow too large relative to the population size. For a large sample size and low rate of recombination, the rate at which the number of ancestors decrease is higher for the discrete model due to an increased incidence of multiple and simultaneous coalescent events.

The effects of increasing the rate of recombination can be seen in the other three plots in Figure 2. The plots corresponding to $R = 1$ and $R = 2.5$ show similar patterns; the behaviors of the continuous and the discrete model are significantly different and the discrep-

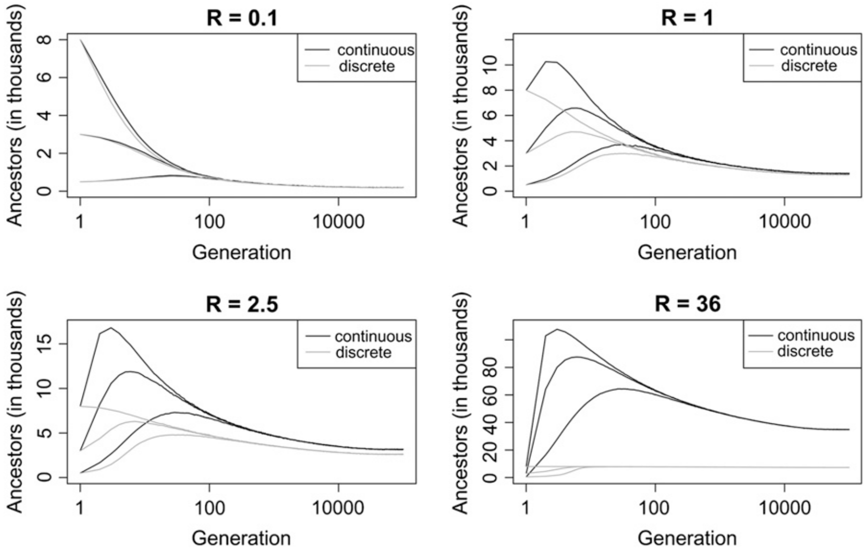

FIGURE 2.—The number of sequences carrying ancestral material as a function of generations back in time. Each plot corresponds to a different recombination rate as labeled.

ancy between them increases as the sample size is increased. In both cases the shape of the approach to the equilibrium and the equilibrium distributions themselves are different. The equilibrium distribution for the discrete model oscillates around a smaller number of ancestors and the peak in the number of ancestors prior to the equilibrium is also smaller. This indicates that the rate of coalescence for the continuous model is significantly underestimated. As R is increased further ($R = 36$) to model the length of the human genome, these effects are amplified.

To distinguish the difference between the equilibrium distributions for the continuous and discrete models, in Figure 3 we plot the average number (estimated over 20,000 generations of a single run of the process) of ancestral sequences as a proportion of the population size once the processes are stationary. For each recombination rate the equilibrium distribution is centered around a larger number of ancestors in the continuous case. As the recombination rate increases, the difference between the equilibrium distributions also gets larger.

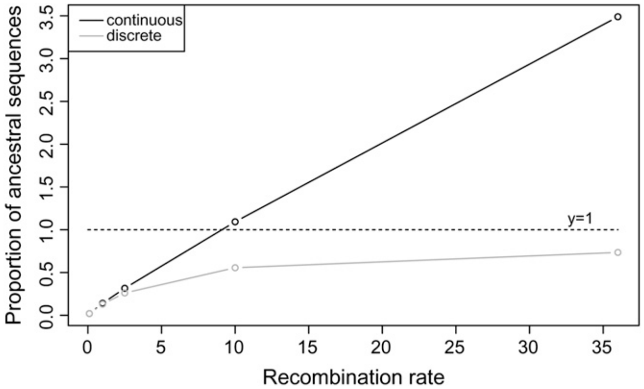

FIGURE 3.—The mean proportion of sequences of the total population that are ancestral to an extant sample (of any size) as a function of recombination rate once the process has reached an equilibrium distribution. The horizontal dashed line is drawn in for reference and shows where this proportion is exactly 1, *i.e.*, where the entire population is ancestral to the sample.

Figure 3 also shows that the mean number of ancestors to which the continuous distribution oscillates around once it has reached the equilibrium can be larger than the effective population size, yielding proportions >1. Although this may seem a strange result, it is consistent with the continuous model since there are no restrictions on the number of sequences that can contain ancestral material. This phenomenon cannot occur with the discrete model since ancestors are chosen from the previous generation that has constant fixed size, thereby imposing $2N$ as an upper bound.

The rate of coalescent events: To estimate how frequently coalescent events occur, we simulate genealogies of a sample of genes and simply count the number of coalescent events. We consider a single coalescent event to be the merging of exactly two sequences in a single generation. Multiple events are counted as the number of sequences coalescing minus one. Simultaneous coalescent events are counted by summing in the natural way. Results are independent of sample size and depend only on recombination rate. For each recombination rate, the expected number of events per generation is estimated from a single run of 20,000 generations taken from the equilibrium distribution. The results are presented in Figure 4 and there is a clear distinction between the discrete and the continuous model. In the continuous case it is possible to calculate exactly the expected number of coalescent events per generation and this illustrated by the dashed line. As Figure 4 shows, our simulations of the continuous model agree with the analytical expectation. The light gray line showing the

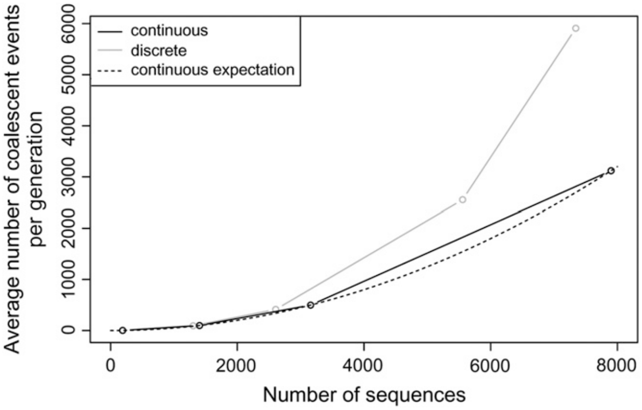

FIGURE 4.—The average number of coalescent events per generation as a function of the number of ancestral sequences. The number of ancestral sequences is specified indirectly by fixing the recombination rate *R*. The intervals at which data points are plotted along the horizontal axis are obtained as the average number of sequences in the equilibrium distributions. The dashed line is the expected number of events calculated exactly from the continuous process.

average number of coalescent events for the discrete model shows the extent to which the continuous model underestimates the rate of coalescence. As the number of ancestral sequences is increased beyond one-fifth of the population size the effect of multiple and simultaneous events becomes significant and raises the rate of coalescence. This confirms our intuition and explains the differences in the equilibrium distributions discussed previously and shown in Figure 2.

The distribution of ancestral material: Figure 2 shows the number of sequences ancestral to the extant sample as a function of time. The number alone gives no indication of how much ancestral material lies on each of the ancestors. Initially the total amount of ancestral material is exactly nR (where n is the sample size). As a function of time, the total amount of ancestral material decreases until each position along the sequence has found a most recent common ancestor. After this point the amount of ancestral material remains constant (namely R) and is redistributed only by subsequent recombination and coalescent events. Our simulations show that the rate at which the total amount of ancestral material decreases does not depend on the model type. In Figure 5 the lines corresponding to the discrete and the continuous model are almost indistinguishable.

We also compare the average total number and average length of ancestral segments once the process describing the number of segments has converged to an equilibrium distribution (Figure 6). Figure 6 (left) shows that the simulation results for the number of ancestral segments at equilibrium agree almost exactly. Furthermore the results are consistent with the exact expectation of the number of segments derived by WIUF and HEIN (1997) for the continuous model (in equilibrium). They derive the expected number of ancestral segments $\mathbb{E}[S] = 1 + \rho/2$ and Figure 6 illustrates this linear relationship.

Our simulation results (Figure 6, right) show that the expected segment length is also approximately the same for the discrete and the continuous model. These results are not surprising since the recombination process that splits the ancestral material into different sequences is modeled in the same way with breakpoints placed at exponential distances with the same expected number of events per generation. The underestimation of the rate of coalescence does not affect this process, unless the extra coalescent events in the discrete model tend to merge segments. Our simulations show that there is no tendency toward the extra coalescent events in the discrete model merging a significant number of segments. Consequently, there are more segments found on a typical ancestor taken from the discrete

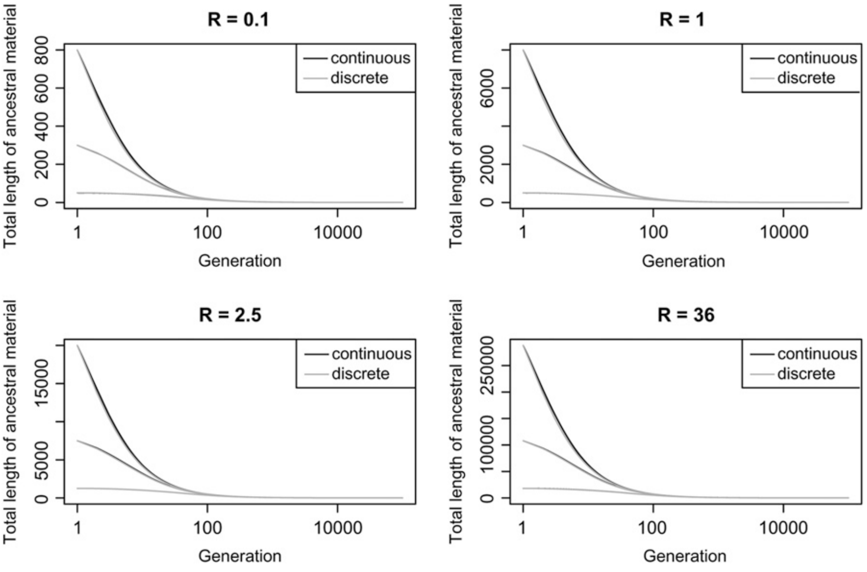

FIGURE 5.—The total amount of ancestral material as a function of generations back in time. Each of the plots corresponds to a different rate of recombination (comparable with Figure 2).

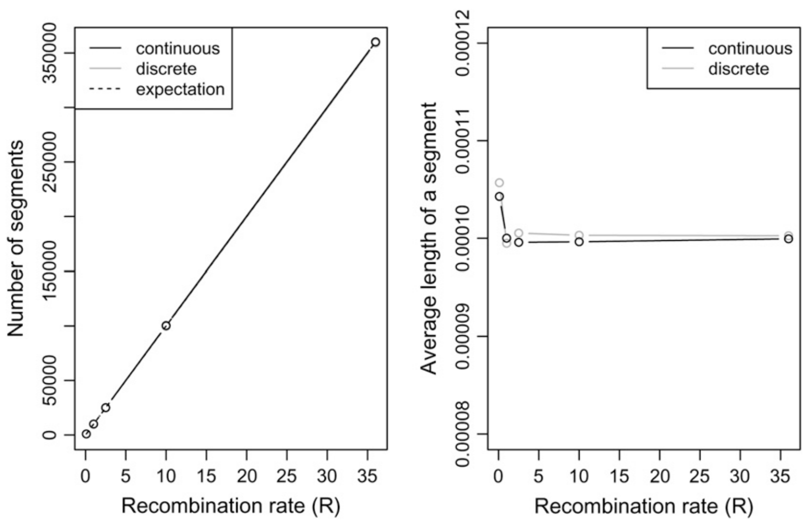

FIGURE 6.—(Left) The average number of segments of ancestral material (once in equilibrium) plotted as a function of recombination rate. (Right) The average segment length as a function of recombination rate.

model compared with a typical ancestor taken from the analogous continuous model.

Shared and correlated ancestries: As described in GRIFFITHS and MARJORAM (1996) the ancestry of a sample of sequences subject to recombination can be described by the Ancestral Recombination Graph (ARG). From the ARG it is possible to trace the ancestry of a single nucleotide or locus by following the appropriate branches in the ARG to produce a coalescent tree. As the distance between two positions increases (and hence the scaled rate of recombination), the correlation of the ancestries decreases. For example, the probability that a recombination event occurs in a small window surrounding a locus is very small and consequently the coalescent trees for the positions contained in this interval are likely to be either identical or very similar to that of the locus itself. Conversely if there is a large distance between two loci it is more likely that recombination events occur between them such that the resulting ancestral histories seem independent with little similarity. There are several measures of correlation and similarity between two trees and such quantities are of great interest with applications in disease association mapping. In particular, we consider the effect of multiple and simultaneous coalescent events on linkage disequilibrium (LD) by Monte Carlo simulation.

There are many qualitative measures of tree similarity defined but many are dependent on the number of leaves of the tree and further, they are not widely used for disease association mapping since knowledge of the two trees corresponding to the loci of interest is required. Instead a measure of correlation of the two loci, r^2, is typically reported. It is defined by

$$r^2 = \frac{(p_{11} - p_1 q_1)^2}{p_1(1 - p_1)q_1(1 - q_1)}, \tag{8}$$

where p_{11} denotes the probability of seeing the wild type in both trees and p_1 and q_1 denote the probability of observing the first and the second wild type, respectively. We simulate the coalescent with recombination under the continuous and the discrete models and construct the coalescent trees for two distinct loci, varying the recombinational distance between them. We choose recombinational distances of 0.1, 2, 10, and 36 and compare the correlation coefficients. The simulations are run until both of the loci have found a most recent common ancestor. The correlation is computed by placing a mutation at random on each of the resulting trees and then considering the proportions of allelic types. For similar trees larger values of r^2 are expected although they are unlikely to yield a value of 1 since this reflects the probability that a mutation is placed on the same branch in the tree.

The mean and the standard deviation for a large sample size of 5000 are displayed in Table 1 and are calculated from 1000 simulations. Simulations from smaller sample sizes display the same pattern. Reassuringly, the average values of r^2 are not significantly different between the models. The reason why r^2 decays at a similar rate can be inferred from the plots in Figure 5.

TABLE 1

Comparison of the approximate mean and variance of r^2 calculated on the basis of 1000 simulations from a sample size of 5000 as the rate of recombination is increased

R	Cont. mean/10^{-4}	Cont. SD/10^{-4}	Disc. mean/10^{-4}	Disc. SD/10^{-4}
0.1	5.195	0.513	4.998	0.476
1	2.394	0.183	2.364	0.188
2.5	2.156	0.155	2.253	0.173
10	2.037	0.135	2.247	0.155
36	2.005	0.132	2.251	0.154

Cont., continuous time; Disc., discrete time.

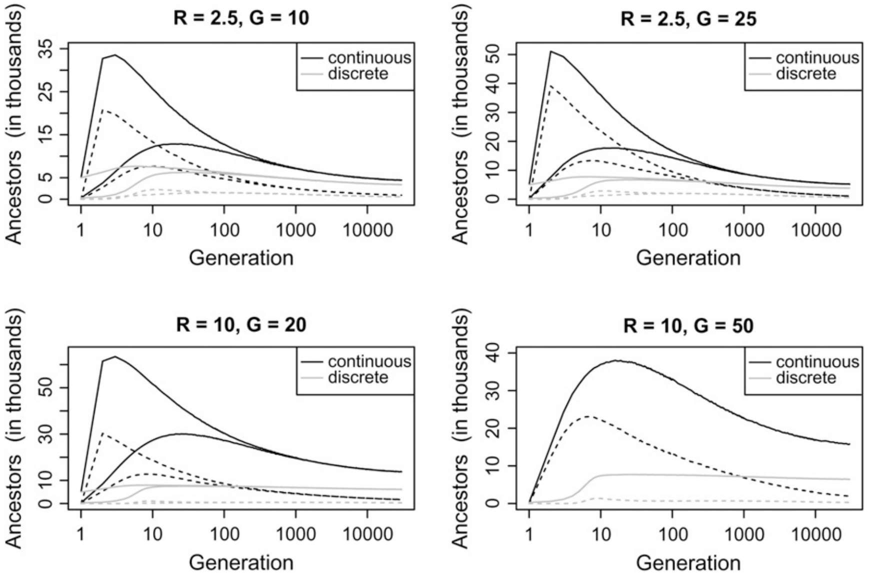

FIGURE 7.—The number of ancestors as a function of generations back in time. Each graph corresponds to fixed rates of recombination (R) and gene conversion (G) as labeled. The additional dashed lines represent the number of ancestors that contain only ancestral material as a result of a gene conversion event. This number is initially zero regardless of the sample size. The larger the sample size is, the higher the peak of the corresponding dashed line.

They show that the rate at which overlapping ancestral material coalesces is the same for both models. Typically the "additional" coalescent events in the discrete model coalesce sequences that have no overlap in the ancestral material they contain. When ancestral material overlaps, the subsets of sample sequences that each of the coalescing sequences is ancestral to have to be disjoint. Without overlap no such restriction exists and we can expect the subsets of sample sequences to which each of the coalescing sequences is ancestral, to be more or less independent of each other. It follows that the value of r^2 decays in a similar way for both models.

The impact of gene conversion: We compare the quantities discussed previously when gene conversion events are also simulated. We vary the ratio of rates (gene conversion to crossover recombination) and use simulation to attain approximate results.

The number of ancestors to a sample: The number of ancestors to an extant sample is a stochastic process as a function of time that converges to an equilibrium distribution. Previously we showed that this equilibrium distribution depends on the rate of recombination and the type of model used, *i.e.*, discrete or continuous. Our simulations including gene conversion yield similar results (Figure 7).

Each gene conversion event places two breakpoints (where possible) on a sequence, whereas a single crossover recombination event places exactly one breakpoint along a sequence. Consequently each gene conversion event generates three segments of ancestral material rather than two segments generated by a crossover recombination event. It follows that adding gene conversion has the effect of raising the overall rate of recombination and our results in Figure 7 confirm this. As the ratio of the rate of gene conversion (G) to the rate of crossover (R) increases we see the proportion of ancestors that contain ancestral material only as a result of a gene conversion event (out of the total number of ancestors) rise. We consider an ancestor to be a gene conversion ancestor only if every segment of ancestral material present on the sequence is a conversion "island," *i.e.*, the segment between the two breakpoints of a single gene conversion event.

The rate of coalescent events: The rate of coalescence is dependent solely on the number of ancestors to the sample; it is not affected by the way in which the ancestral material is distributed onto ancestral sequences. Increasing the rate of gene conversion increases the overall rate at which recombination events (crossover and conversion) occur and hence raises the average equilibrium number of ancestors. The rate of coalescence given the number of sequences remains the same for the discrete and the continuous model.

The distribution of ancestral material: The length of a segment of ancestral material depends on whether the segment was generated by a gene conversion event or by a crossover event. Segments cut out by a gene conversion event are of fixed length $(R + G) \times 10^{-6}$ and are rarely subsequently hit by another crossover or conversion event because they are small. This length is the same in the continuous model and the discrete model. The main difference we see in the distribution of ancestral material on a typical ancestor with the addition of gene conversion is that often there are gene conversion islands taken out of the recombination crossover segments.

The total number of ancestral segments with the addition of gene conversions is also independent of the model choice. Each crossover event creates one breakpoint and one extra segment, whereas each gene conversion event creates two breakpoints and two extra segments. For consistency with our results in Figure 6 we expect the average total number of segments to be $1 + \rho/2 + \gamma$ (since a conversion event generates twice as many new segments as a crossover event). Our simulations confirm this for both models. The independence of model choice is also as expected since the underestimation of the rate of coalescence changes the average number of ancestors to a sample and the number of segments found on each ancestor, but not the total number of ancestral segments.

DISCUSSION

Our results show that of the quantities we considered, only the average rate of coalescence and the equilibrium average number of ancestors to a sample (of any size) are affected by the use of Hudson's continuous-time approximation. In particular, linkage disequilibrium is not affected. This result is not trivial or immediately intuitive.

The larger average equilibrium number of ancestors displayed by the continuous model is explained by the underestimation of the rate of coalescence. The additional multiple and simultaneous coalescent events that occur in the discrete simulations typically place disjoint nonoverlapping regions of ancestral material onto the same ancestor. Hence the rate at which the total amount of ancestral material decays remains the same for both models, and on average the number of segments on an ancestor (in equilibrium) is greater for the discrete model.

It is not obvious that r^2 should decay with similar rates for the two models, particularly because the initial behavior of models and their equilibrium distributions do not agree for large recombination rates. With fewer ancestral sequences, one might expect r^2 to decay less rapidly in the discrete model. However, the rate at which overlapping ancestral material coalesces is the same for both models and when two or more segments of ancestral material are placed onto the same ancestor in the discrete model they are usually separated by a large region of nonancestral material. So there are two possible factors that result in the same decay of r^2. First, the time prior to two segments being placed on the same ancestral sequence allows for these segments to have very different histories. Second, ancestral segments may again be split onto different ancestors when recombination events put an odd number of breakpoints between the two segments. So despite events being shared between the segments for a period of time, this leaves little or no trace as the segments may carry ancestral material for unrelated subsets of the sample and the segments may again become separated in their further ancestry.

The quantities we calculate from our simulations can be considered in two classes: global or local. Global quantities are calculated with knowledge of the ancestral material present on each entire sequence and include the total number of ancestral sequences and the total rate at which coalescent events occur (since these require knowledge about whether any region of the sequence is ancestral). Local quantities are calculated on a region/segment level and include the number of segments of ancestral material, segment length, and the total amount of ancestral material. Also r^2 can be considered local in the sense that it is computed with knowledge only of the ancestry of the end points of the sequence, not that of the entire sequence.

Our simulation results show that the global quantities are affected by the additional multiple and simultaneous coalescent events in the discrete-time model, while the local quantities including r^2 are not. The total number of sequences ancestral to the sample is overestimated by the continuous model and the error margin increases as the rate of recombination increases. The rate of coalescence is vastly underestimated by the continuous model, as expected, and the increase of the error margin of the total number of ancestral sequences is a reflection of the extent to which the rate of coalescence is underestimated with larger recombination rates.

Local quantities regarding ancestral segments and linkage disequilibrium are indistinguishable between models. The total number and the length of segments are determined by the recombination process rather than the coalescent process and therefore are not affected by the underestimation of the rate of coalescence in the continuous model. The rate at which the total amount of ancestral material decays is also indistinguishable between the continuous and the discrete model. This is an important result for disease association mapping since the knock on effect is that r^2 is not affected by the use of the continuous-time approximation.

We thank Carol Teo for her collaboration on an introductory summer project with Joanna Davies.

LITERATURE CITED

Fu, Y. X., 2006 Exact coalescence for the Wright-Fisher model. Theor. Popul. Biol. **69:** 385–394.

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

Hein, J., M. H. Schierup and C. Wiuf, 2005 *Gene Genealogies Variation and Evolution.* Oxford University Press, London/New York/Oxford.

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

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

Kong, A., D. F. Gudbjartsson, J. Sainz, G. M. Jonsdottir, S. A. Gudjonsson *et al.*, 2002 A high resolution recombination map of the human genome. Nat. Genet. **31:** 241–247.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. **27:** 1870–1902.

Sagitov, S., 1999 Coalescents with simultaneous multiple collisions. Electron. J. Probab. **36:** 1116–1125.

Sagitov, S., 2003 Convergence to the coalescent with simultaneous multiple mergers. J. Appl. Probab. **40:** 839–854.

Schweinsberg, J., 2000 Coalescents with simultaneous and multiple collisions. Electron. J. Probab. **5:** 1–50.

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

Communicating editor: Y.-X. Fu