

# On Recombination Induced Multiple Coalescent Events

**Joanna Davies and František Šimančík**  
Worcester College



Oxford Centre for Gene Function  
Department of Statistics  
University of Oxford

**September 2006**

**Abstract**

# Contents

<b>1</b>	<b>Introduction and Motivation</b>	<b>2</b>
<b>2</b>	<b>Background</b>	<b>2</b>
2.1	The Wright-Fisher Model with Recombination . . . . .	2
2.2	Hudson's Continuous Time Model . . . . .	3
2.2.1	The rate at which sequences are involved in a recombination event and a coalescent event simultaneously . . . . .	3
2.2.2	Derivation of the waiting time until a recombination event . . . . .	4
2.2.3	Derivation of the waiting time until a coalescent event . . . . .	4
2.2.4	Simulating from the continuous model . . . . .	5
2.3	Gene Conversion . . . . .	5
2.3.1	Simulating gene conversion with the discrete Wright-Fisher model . . . . .	6
2.3.2	Simulating gene conversion with Hudson's continuous time model . . . . .	6
<b>3</b>	<b>Methodology and Results</b>	<b>7</b>
3.1	The Number of Ancestors to a Sample . . . . .	7
3.2	The Rate of Coalescent Events . . . . .	9
3.3	The Distribution of Ancestral Material . . . . .	10
3.4	Ancestral Material on a Typical Common Ancestor . . . . .	12
3.5	Shared and Correlated Ancestries . . . . .	13
3.6	The Impact of Gene Conversion . . . . .	15
3.6.1	The number of ancestors to a sample . . . . .	15
3.6.2	The rate of coalescent events . . . . .	16
3.6.3	The distribution of ancestral material . . . . .	16
3.6.4	Ancestral Material on a Typical Common Ancestor . . . . .	17
<b>4</b>	<b>Conclusion</b>	<b>19</b>

# 1 Introduction and Motivation

The stochastic process modelling the coalescent with recombination on a continuous time scale is well described by Hudson [1]. The derivation of the continuous time process is obtained by rescaling the discrete time Wright Fisher model and taking the limit as the population size tends to infinity. This derivation is discussed in section 2 but is only a valid approximation provided that the sample size of extant sequences is small relative to the effective population size. In particular it is assumed that going back in time, multiple and simultaneous coalescent events do not occur, and further that ancestral sequences always recombine with (infinitely many) non-ancestral sequences and never with each other. As the sample size increases relative to the population size the probability of such events occurring becomes non-negligible and in such cases it is interesting to investigate the effect of these events on the rate of coalescence and other quantities typically computed.

It may be argued that any large sample under Hudson's model will quickly reduce down to a small sample such that the assumptions soon become valid; however, 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.

## 2 Background

### 2.1 The Wright-Fisher Model with Recombination

A coalescent model is typically used to infer details about the evolutionary processes yielding a sample of extant DNA sequences from a population. Coalescent processes can be viewed either forwards or backwards in time and can accommodate many factors including population growth, recombination, migration and selection.

The most basic coalescent model is the Wright-Fisher model, which can be used to describe the genealogy of a haploid population of constant size. Genealogies of a sample from the population can be simulated back in time; for each individual in the current generation a parent is chosen at random from the previous generation. Equivalently forwards in time, the next generation is simulated by selecting each individual in the new generation at random (with replacement) from the previous generation.

The basic model uses the following assumptions to simplify the process.

1. All individuals in the population are haploid.
2. All individuals are equally fit.
3. The population remains constant.
4. Time is measured in discrete non-overlapping generations.
5. The population has no geographical or social structure.
6. The sequences in the population are not recombining.

In practise assumptions 1 and 4 although invalid are of little practical consequence. Extensions can be made to the basic model as discussed in [2] chapter 4 to account for 3 and 5. Incorporating selection (relaxing assumption 2) is discussed in [15]. Perhaps the most important assumption to be relaxed is 6. DNA sequences are almost always subject to recombination. The extension to the Wright-Fisher model to include recombination is also described in [2] but the details and set up is described here since it is from this model we simulate genealogies to investigate the effects of non-standard coalescent and recombination events.

The genealogy of the genetic material of a sample of  $n$  individual sequences is to be simulated using the Wright Fisher model with recombination. Let the population of haploid individuals be of constant size  $2N$  with sequence length  $L$  nucleotides. To simulate the distribution of genetic material ancestral to the extant sample in the previous generation we use the following setup.

Let  $r$  be the probability of a recombination between two nucleotides in a sequence per generation. Let the length of each sequence be  $L$  nucleotides such that the expected number of recombination events is  $rL = R$ . Since  $r$  is

assumed to be very small and  $L$  large, the binomial distribution on the number of recombination events in a single sequence per generation is well approximated by a poisson distribution with mean  $R$ . Hence the recombination events on a single sequence of scaled length  $R$  in a single generation can be modelled by a poisson process with constant rate 1. The parameter  $R$  is known as the scaled sequence length and in practise we tie  $r$  and  $L$  together by specification of  $R$  directly. This allows  $R$  to be well defined in the limit as  $r \rightarrow 0$  and  $L \rightarrow \infty$  if desired. Then the resulting evolutionary process can be simulated back in time in the following way.

1. Start with  $k = n$  individual sequences. Initially the entirety of each sequence is ancestral material.
2. For each of the  $k$  sequences carrying ancestral material, choose two parents at random from the population in the previous generation. Note that it is possible to choose the same parent twice. In this instance no recombination occurs and all the genetic material is transferred to the single parent.
3. Place recombination events along the current sequence of length  $R$  according to a poisson process with rate 1 such that genetic material between or on either side of the recombination events is distributed alternately on each of the two parents. In practise this can be done by drawing exponential distances between recombination events with parameter 1. Note that if the first distance simulated is greater than  $R$  then no recombination event occurs and one of the parents selected at random receives all the ancestral material.
4. Update  $k$  to be the new number of ancestral sequences after steps 2 and 3 and go back to 2.

Simulating recombination events in this way automatically simulates the coalescent events which occur when two or more sequences containing ancestral material choose the same parent(s) in the previous generation. A multiple coalescent event occurs when more than two sequences choose the same parent. The nature of the process also allows two or more single or multiple coalescent events to occur simultaneously in a generation. Step 2. allows an individual to choose the same parent twice. This is unrealistic but the probability this happens is  $1/10000$  and it rarely happens in the simulations in practise. Also note that 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  $k$  extant sequences evolves according to the basic coalescent without recombination and the resulting coalescent tree is local to that position. Consequently the genealogy relating a set of  $k$  extant sequences with recombination can be considered a collection of local trees. [9]

## 2.2 Hudson's Continuous Time Model

Hudson [1] derives the continuous time coalescent model with recombination from 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.

### 2.2.1 The rate at which sequences are involved in a recombination event and a coalescent event simultaneously

The probability that a sequence is involved in both a recombination and a coalescent event simultaneously is the probability of the sequence being involved in a recombination event multiplied by the probability that the sequence coalesces with another since recombination events and coalescent events occur independently. In discrete time when  $R$  is small and  $N$  is large this quantity is given by (1).

$$\frac{1}{2N} \times R = \frac{1}{2N} \times \frac{\rho}{4N} = \frac{\rho}{8N^2} \quad (1)$$

For fixed  $\rho$  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  $\rho$  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.

### 2.2.2 Derivation of the waiting time until a recombination event

Hudson [1] derives the continuous time coalescent model with recombination from the discrete time Wright-Fisher model with recombination. Using the set up above for the discrete model, consider the waiting time (denoted  $T$ ) 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 as specified by (2) for  $m \in \mathbb{N}$

$$P(T = m) = (e^{-R})^{m-1}(1 - e^{-R}) \quad (2)$$

To take the continuous limit, let  $N \rightarrow \infty$  and  $R \rightarrow 0$  and define  $\frac{\rho}{2} := 2NR$ . Since  $R$  is small,  $e^{-R} \approx 1 - R$  such that

$$P(T = m) = R(1 - R)^{m-1}$$

Hence when time is scaled to be measured in  $2N$  generations denoting the continuous waiting time by  $T_C$ ,

$$P(T_C \leq t) = 1 - (1 - R)^{[2Nt]} = 1 - \left(1 - \frac{2NR}{2N}\right)^{[2Nt]} = 1 - \left(1 - \frac{\rho}{4N}\right)^{[2Nt]} \approx 1 - e^{-\frac{\rho t}{2}} \quad (3)$$

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 break point is placed uniformly along the selected sequence.

### 2.2.3 Derivation of the waiting time until a coalescent event

The rate at which coalescent events occur in the Wright-Fisher model with recombination is the same as that of the basic model. A coalescent event occurs between any two sequences in a single generation with probability  $\frac{1}{2N}$  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 (4).

$$\prod_{j=1}^k \frac{(2N - (j - 1))}{2N} = \prod_{j=1}^{k-1} \left(1 - \frac{j}{2N}\right) \quad (4)$$

This product can be expanded (5). When the sample size  $k$  is small relative to the population size  $2N$  the term  $O\left(\frac{1}{N^2}\right)$  (gathering all remaining terms in the expansion) is negligible.

$$\prod_{j=1}^{k-1} \left(1 - \frac{j}{2N}\right) = 1 - \sum_{j=1}^{k-1} \frac{j}{2N} + O\left(\frac{1}{N^2}\right) = 1 - \binom{k}{2} \frac{1}{2N} + O\left(\frac{1}{N^2}\right) \quad (5)$$

Under such conditions, the probability that no coalescent event occurs in a single generation is (6).

$$1 - \binom{k}{2} \frac{1}{2N}, \quad (6)$$

The probability of a single coalescent event occurring in a single generation is given by (7).

$$\binom{k}{2} \frac{1}{2N} \quad (7)$$

The term  $O(1/N^2)$  is the term which sums the probabilities of all possible multiple and simultaneous coalescent events. However, this summation of terms is only negligible and of order  $1/N^2$  provided that  $k \ll 2N$ . As  $k$  approaches  $2N$  these terms get larger and cannot be neglected.

The derivation of the continuous time coalescent model for a sample of  $k$  sequences from a population of size  $2N$  first derived by Kingman (1982) [6][7], is based on this approximation and the assumption that  $k \ll 2N$ . Under the approximation, the waiting time while there are  $k$  ancestors (denoted  $T_k$ ), is geometrically distributed with mean  $2N \binom{k}{2}^{-1}$ . The continuous time coalescent process is obtained by scaling time to be measured  $2N$  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 (denoted  $T_C$ ) 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 are chosen at random out of the possible  $\binom{k}{2}$ .

#### 2.2.4 Simulating from the continuous model

The ancestry of a small sample of sequences can be determined according to the coalescent and recombination events occurring back in time. A recombination event increases the number of sequences with ancestral material by 1 and a coalescent event decreases the number of sequences by one and since the waiting times to each of these events are independent and exponentially distributed, they can be viewed as competing processes. Hence under the continuous model of the coalescent with recombination it is possible to simulate the ancestry of  $k$  sequences using the following method also described in [2].

1. Start with  $k = n$  sequences of length  $R$ .
2. Simulate the time until a coalescent or recombination event occurs within the  $k$  sequences by drawing a time from an exponentially distributed random variable with parameter  $\binom{k}{2} + k\rho/2$ .
3. Determine whether the event is to be a recombination or a coalescence. With probability  $(k-1)/(\rho+k-1)$  it is a coalescence, otherwise it is a recombination.
4. If the event is a recombination, place the recombination break point uniformly along the sequence. If there is ancestral material either side of the breakpoint, create two ancestor sequences one containing the ancestral material to the left of the break point and the other containing the ancestral material to the right. Increase the number of sequences containing ancestral material i.e set  $k \leftarrow k + 1$  by one and go back to 2. If the breakpoint does not divide the ancestral material create an ancestor which is passed all the ancestral material and leave the number of sequences containing ancestral material unchanged.
5. If the event is a coalescence, choose two sequences from the  $k$  to coalesce. Create a single ancestor for the two sequences by creating an ancestor which is passed the ancestral material from both of the coalescing sequences. Update  $k \leftarrow k - 1$  and go back to 2.

Just as with the discrete Wright-Fisher model with recombination, ancestral material of a single sequence can be split onto many sequences such that potentially each position may have a different coalescent tree. Each of these trees can be embedded into the Ancestral Recombination Graph (ARG) first introduced by Griffiths and Marjoram [9]. The simulation described above constructs exactly an ARG, that is the entire genealogy of the sample. We can also construct the ancestral recombination graph for the discrete time Wright-Fisher model with recombination via the simulation method described in section 2.1.

### 2.3 Gene Conversion

The algorithms discussed previously to simulate the ancestry of a sample of sequences place recombination breakpoints along sequences at exponential distances. These breakpoints typically 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 DNA from one chromosome to another. They are called homologous gene conversion events and are thought to occur more frequently than would be expected by drawing a very small distance from exponential distribution. They can be modelled directly by adding a rate of gene conversion. Gene conversion events occur independently of recombination and coalescent events and they can be simulated in a similar way. The length of the small fragment to be transferred can either be fixed or taken from another distribution. In the human genome fragments lengths vary between 100 and 300 kilobases which is approximately one millionth of total length of ancestral material. 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.

### 2.3.1 Simulating gene conversion with the discrete Wright-Fisher model

1. Start with  $k = n$  sequences each of length  $G + R$ .
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 and continue to the next sequence. If two distinct parents are chosen, place gene conversion and recombination events along the sequence by proceeding to 3.
3. Simulate the distance along the sequence at which an event occurs by simulating from an exponential random variable with parameter 1. Where possible, place a breakpoint along the sequence at this distance (from the previous one if applicable). If the distance stretches beyond the length of the sequence, go straight to 6.
4. With probability  $G/(G + R)$  it is a gene conversion event and with probability  $R/(G + R)$  it is a recombination event. If it is a gene conversion event go to 5, otherwise record the breakpoint and go back to 3.
5. Place another breakpoint on the sequence at distance one millionth upstream of the first breakpoint where this is possible. If it is possible (i.e. one millionth following the breakpoint doesn't stretch beyond the length of the sequence) go back to 3. Otherwise go to 6.
6. Distribute the ancestral material between breakpoints alternately onto the two parents chosen at random from step 2. Check if all sequences have been considered. If not simulate recombination and gene conversion events for the next sequence. Otherwise go to 7.
7. Update  $k$ , the current number of ancestral sequences and go back to 2.

### 2.3.2 Simulating gene conversion with Hudson's continuous time model

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  $\binom{k}{2} + k\rho/2 + k\nu/2$ . ( $\nu := 4NG$ )
3. Determine the type of event. With probability  $k - 1/(k - 1 + \rho + \nu)$  it is a coalescent event, with probability  $\rho/(k - 1 + \rho + \nu)$  it is a recombination (crossover) event and with probability  $\nu/(k - 1 + \rho + \nu)$  it is a gene conversion event.
4. If it is a recombination or a coalescent event proceed in same way described previously. Update the current number of ancestors to the sample and continuing. Otherwise if it is a gene conversion event place a breakpoint along the sequence uniformly go to 5.
5. Place another breakpoint along the sequence at a distance one millionth up from the initial one. The distribute the ancestral material on two newly created ancestors. Place the ancestral material between the two break points onto one of the ancestors and the remainder on the other. Update the current number of ancestral sequences i.e. set  $k \leftarrow k + 1$  and go back to 2.

Modelling 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 only choose 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. We look at features of an ancestry to a sample for the continuous and discrete models with and with without gene conversion in the following section.

### 3 Methodology and Results

We investigate the effect of multiple and simultaneous coalescent/recombination events on features of the Ancestral Recombination Graph via Monte Carlo simulation. There are many interesting quantities which can be extracted from the simulations and directly compared. In particular, Wiuf and Hein ([3], [5]) discuss the number of most recent common ancestors along the sequence to a sample (as a proportion of the population size), the time taken until each position on the sequence finds a most recent common ancestor, the number of segments of ancestral material distributed on these ancestors and the average segment length. Also of interest is the correlation between genealogies at different sites. Here we explain how all of these features are affected by multiple and simultaneous coalescent/recombination events.

#### 3.1 The Number of Ancestors to a Sample

The number of sequences carrying ancestral material (i.e. the number of ancestors) at time  $t$  is a stochastic process. It is described by a Markov chain (either continuous or discrete) and these processes can be shown to converge in both the discrete and continuous case to an equilibrium distribution (see [14]) which is independent of the initial sample size and dependent only upon the rate of recombination. This is demonstrated by figure 1.

Each of the plots in figure 1 correspond to simulations run with different recombination rates (as indicated in the figure). Various sample sizes for both the continuous and discrete model are plotted in black and red 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 1) 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 (namely 8000 out of a population of 10000) the rate at which the number of ancestors decrease is higher for the discrete model due to increased incidence of multiple and simultaneous coalescent events.

The effects of increasing the rate of recombination can be seen by the other three plots. The values of  $R$  we use correspond to the human genome. The value  $R = 2.5$  is approximately the rate of recombination along a single chromosome and the value  $R = 36$  that of the entire genome. The plots corresponding to  $R = 2.5$  and  $R = 1$  show similar patterns; the behaviour of the continuous and the discrete model are significantly different and the discrepancy 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 is not surprising and indicates that the rate of coalescence for the continuous model is significantly underestimated. We investigate this further in section 3.2. As  $R$  is increased further ( $R = 36$ ) to the human genome, these effects are amplified.

To distinguish the difference between the equilibrium distributions for the continuous and discrete models 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 oscillates around a larger number of ancestors in the continuous case. As the recombination rate increases, figure 2 shows that the difference between the equilibrium distributions also gets larger. The discrepancy is explained by the underestimation of the rate of coalescence in the continuous case (caused by neglecting the possibility of simultaneous and multiple events). We continue to investigate how frequently non-standard coalescent events occur in section 3.2.

**Figure 1** The number of sequences carrying ancestral material as a function of generations back in time. Each plot corresponds to a different recombination rate as labelled.

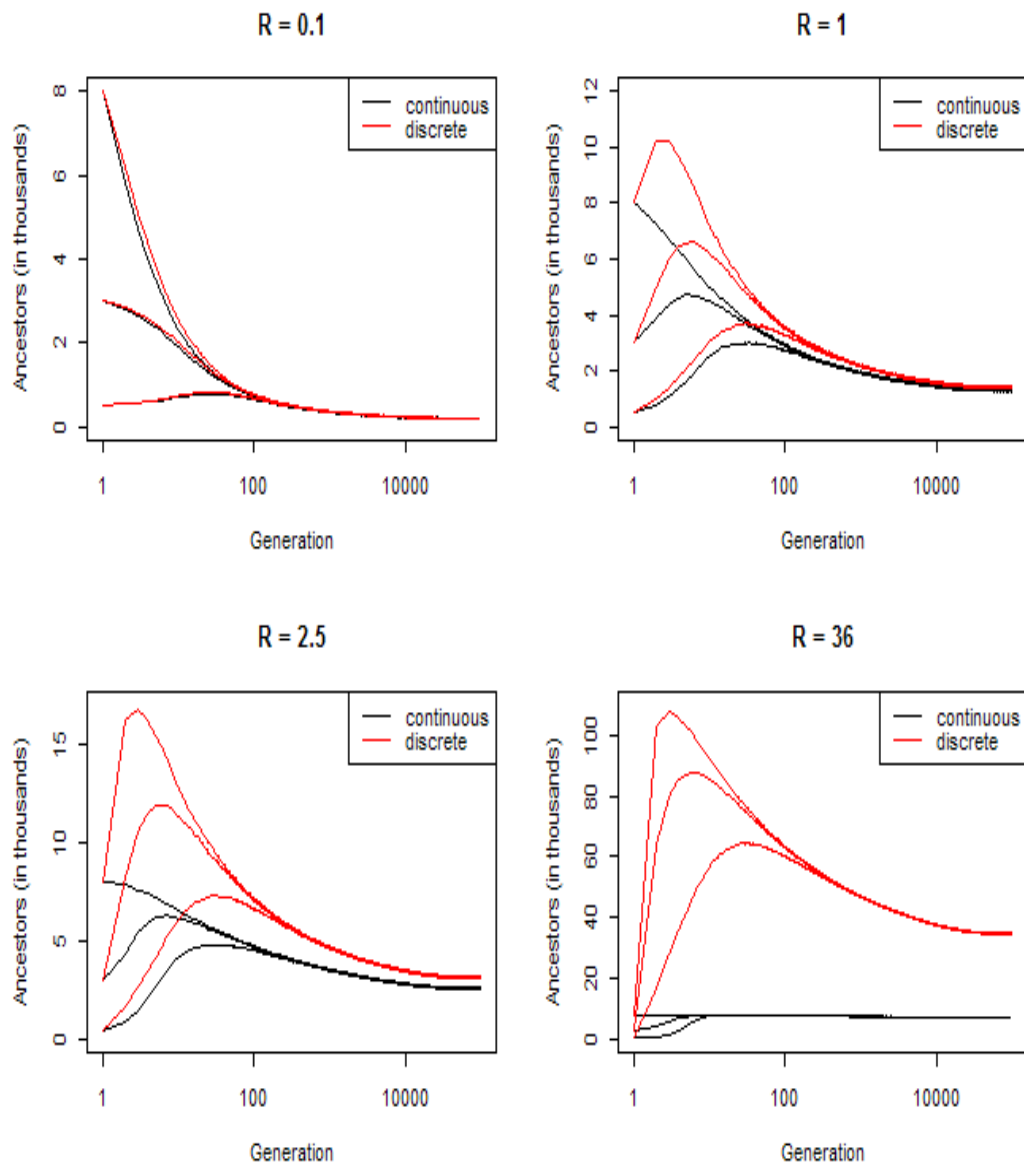
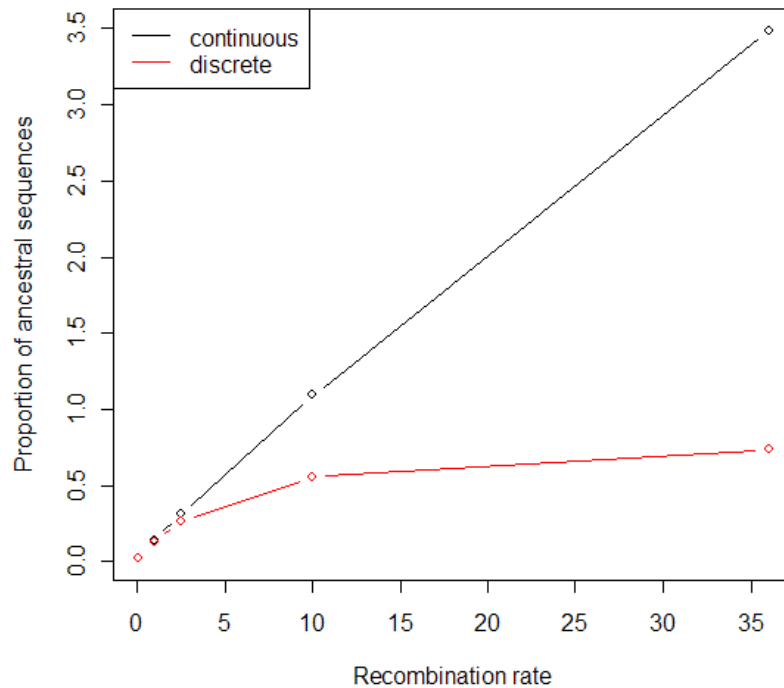


Figure 2 also shows that the mean number of ancestors to which the continuous distribution oscillates around once it has reached the equilibrium can be larger than the effective population size, yielding proportions greater than 1. Although this may seem a strange result, it is consistent with the continuous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

---

**Figure 2** The mean number of ancestors to a extant sample (of any size) as a function of recombination rate once the process has reached an equilibrium distribution

---

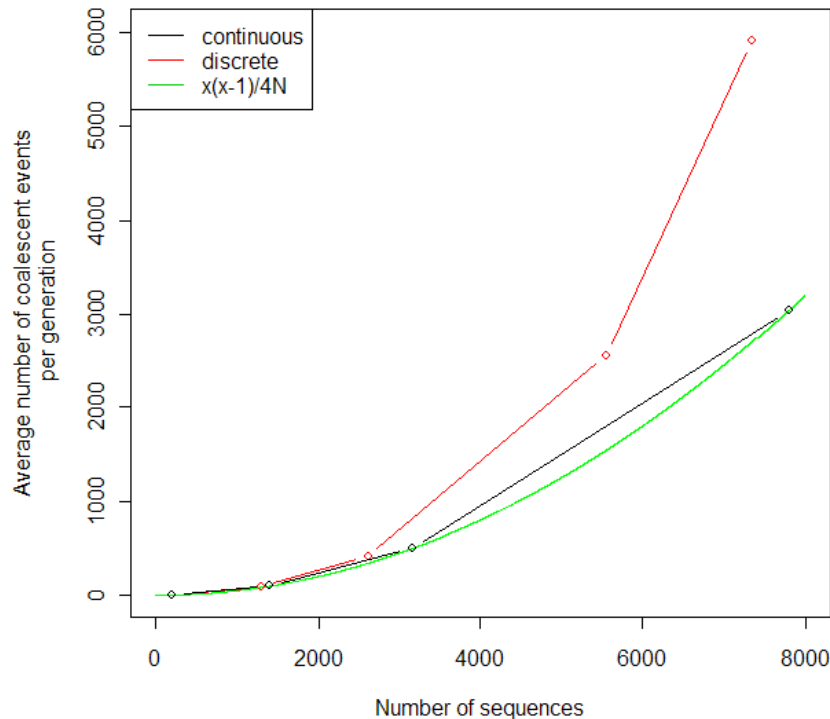


---

### 3.2 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. The results presented are for a population size of 10,000. We wait until the process describing the number of ancestors converges to the equilibrium distribution before starting the counts, so the 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 3 and there is a clear distinction between the discrete and continuous model. In the continuous case it is possible to calculate exactly the expected number of coalescent events per generation and we plot this number in green. As figure 3 shows, our simulations of the continuous model agree with the analytical expectation. The red line showing the average number of coalescent events for the discrete model shows the extent to which the continuous model is underestimating the rate of coalescence. As the number of ancestral sequences is increased beyond 1/5th 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 1.

**Figure 3** The average number of coalescent events per generation as a function of the number of ancestral sequences. The number of ancestral sequences is dependent on the rate of recombination ( $R$ ) and in practise, each point along the horizontal axis corresponds simulations using a different recombination rate (in equilibrium). The red line show the simulation results for the discrete model and the black line for the continuous model. The green line is the expected number of events calculated exactly from the continuous process.



### 3.3 The Distribution of Ancestral Material

Figure 1 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). Recombination events redistribute the ancestral material without changing the total amount (provided there is no recombination with a sequence already ancestral to the sample). Coalescent events reduce the amount of ancestral material being traced if the two sequences chosen to coalesce have ancestral material in common locations. Consequently 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 only redistributed 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 type of model. In figure 4 the lines corresponding to the discrete and continuous model are almost indistinguishable.

The underestimation of the rate of coalescence in the continuous model makes this result counterintuitive and we expected the decay of the total amount of ancestral material to be faster for the discrete model. This is because the majority of ‘additional’ coalescent events involve the coalescence of disjoint pieces of ancestral material. Consequently the number of ancestors to the sample is in general smaller although the total amount of ancestral material remains the same.

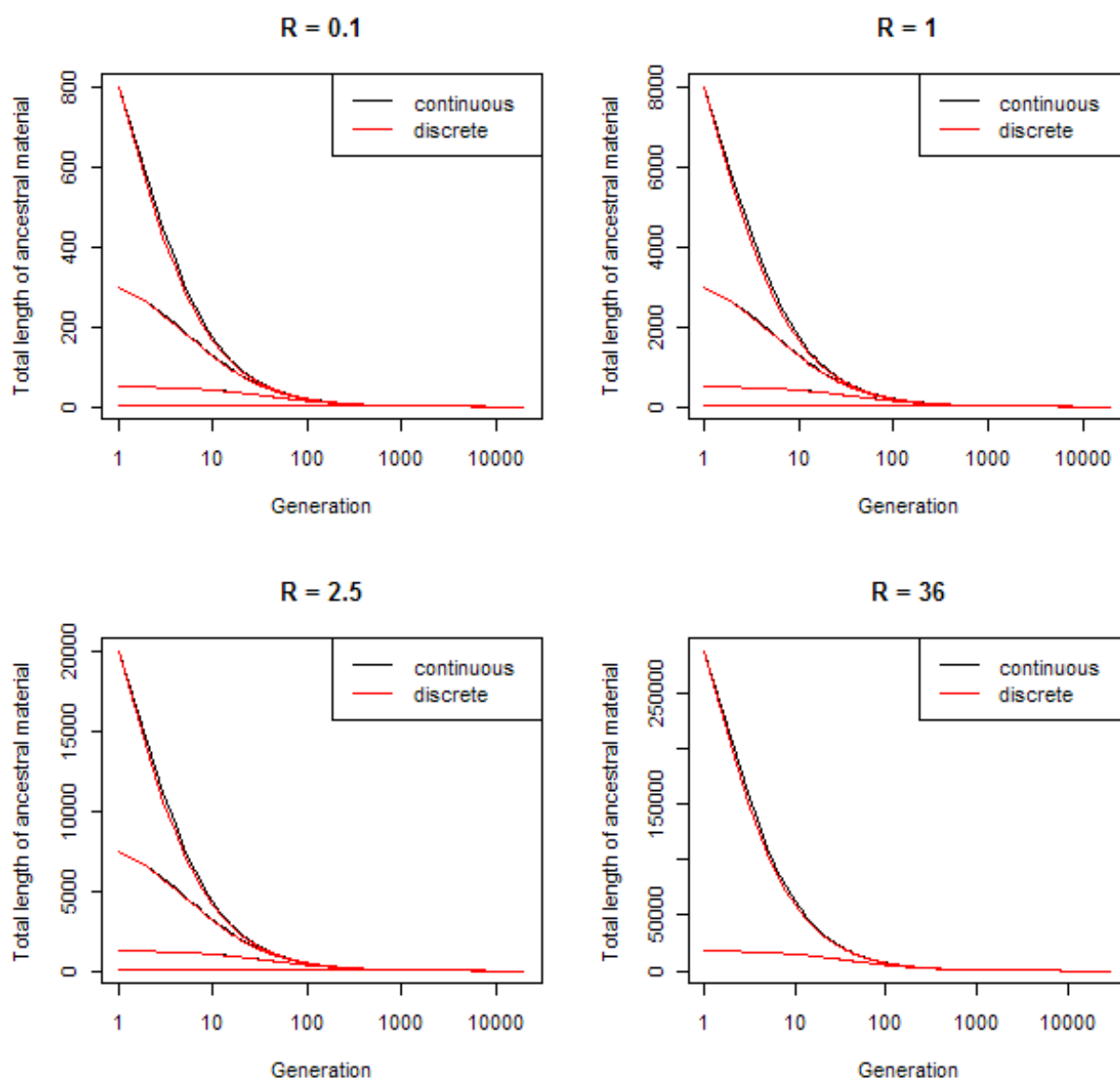
We also compare the average total number and average length of ancestral segments once the process describing the number of segments has converged to an equilibrium distribution (figure 5). Our simulations are taken from the first 20,000 generations after the time that the total amount of ancestral material on all ancestors sums to  $R$ . At this point the process describing the total number of ancestors to the sample has converged to an equilibrium distribution. We recognise that this need not be the case for example if we simulate the ancestry of a single

sequence at time zero all positions have a common ancestor but for our simulations since we use larger sample sizes, taking 20,000 generations after this time ensures that the process is in equilibrium.

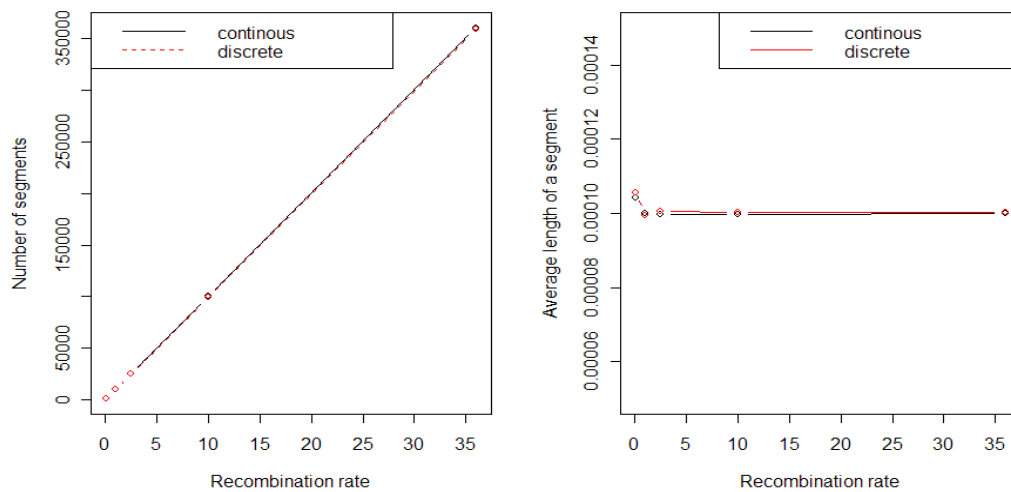
Figure 5 (left) shows that the simulation results for the number of ancestral segments at equilibrium agree almost exactly. Furthermore the results are consistent with the exact expectation of the number of segments derived by Wiuf and Hein [3] for the continuous model (in equilibrium). They derived the expected number of ancestral segments ( $\mathbb{E}[S]$ ) given by (8) and the plot as a function of recombination rate is the straight line we see from the simulations in 5.

$$\mathbb{E}[S] = 1 + \rho/2 \tag{8}$$

**Figure 4** The total amount of ancestral material as a function of generations back in time. Each of the plots correspond to a different rate of recombination (comparable with with figure 1) and the black and red lines represent results from the continuous and discrete model simulations respectively.



**Figure 5** (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 length.



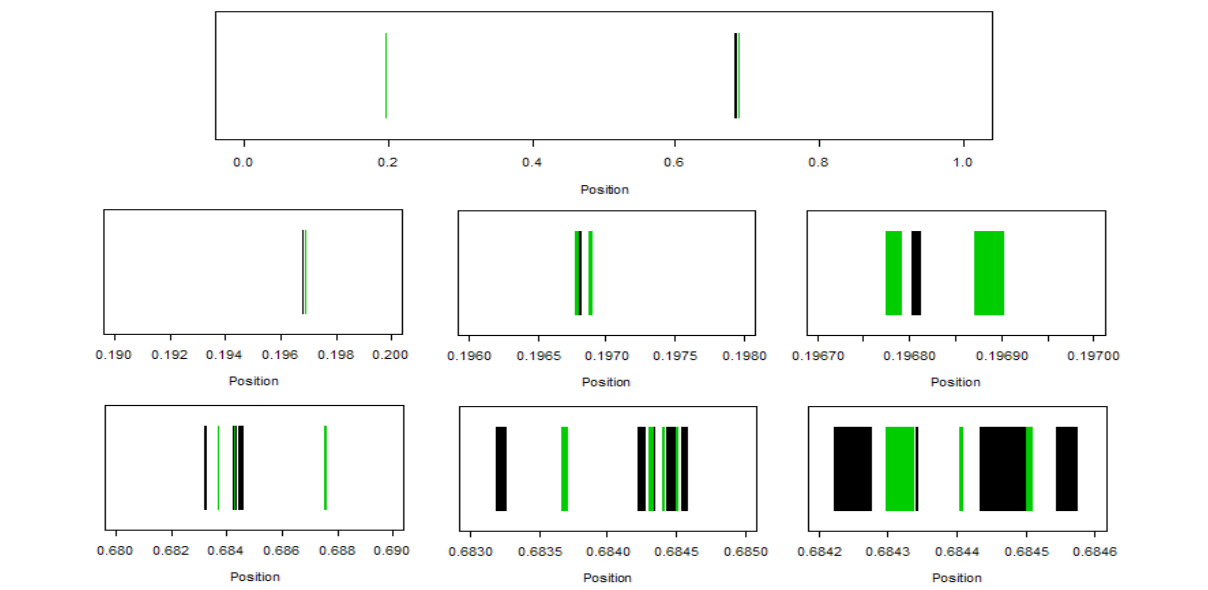
Our simulation results (figure 5 (right)) show that the expected segment length is also approximately the same for the discrete and continuous model. These results are not surprising since the recombination process which spits the ancestral material onto different sequences is modelled in the same way 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, it affects the number ancestors on which these segments are distributed. There are more segments found on a typical ancestor taken from the discrete model compared with a typical ancestor taken from the analogous continuous model. We proceed to investigate the distribution of ancestral material on a single ‘typical’ common ancestor.

### 3.4 Ancestral Material on a Typical Common Ancestor

The recombination process redistributing segments of ancestral material is essentially the same for both the continuous and the discrete model. Consequently, as demonstrated in the previous section, the total number of segments and total amount of ancestral material is consistent across the models. It is interesting to compare the distribution of ancestral material on a typical ancestor once the process describing the number of ancestors has reached an equilibrium distribution. As expected the distribution is similar for both models (figure 6) although the number of regions containing segments of ancestral material is greater for a typical ancestor from the discrete model.

At the moment we only have one model. We need to insert another to compare the models.

**Figure 6** The distribution of ancestral material on an ancestor chosen at random from the discrete process with  $R = 2.5$ . The top picture displays the whole sequence and the subsequent pictures are of the two regions containing ancestral material on a larger scale. The segments of ancestral material are coloured green and black alternately to highlight where recombination events have occurred.



### 3.5 Shared and Correlated Ancestries

As described in [9] 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 it is of great interest with applications in disease association mapping. We consider the effect of multiple and simultaneous coalescent events on Linkage Disequilibrium (LD) and other measures of tree similarity by Monte-Carlo simulation.

The qualitative measures of tree similarity are described in [2]. The first is defined to be the probability that a mutation in each of the trees would create the same bi-partition of the sequences into ancestral and mutant types. This is the probability that mutations on the two different trees create the same sample configuration. The mutations are placed at random uniformly on the trees. To compute the measure it is necessary to determine pairs of branches (one in each tree) which yield identical bi-partitions, then the sum defining the measure of similarity is taken over all these possible pairs. More specifically for trees  $A$  and  $B$  with total branch lengths  $l_A$  and  $l_B$  respectively, the measure is defined by (9). The indicator function  $I_{\{i=j\}}$  is 1 if a mutation on branch  $i$  in tree  $A$  yields the same bi-partition as a mutation on branch  $j$  in tree  $B$  and zero otherwise. The length of branches  $i$  and  $j$  are denoted by  $a_i$  and  $b_j$  respectively.

$$M_{AB} = \frac{1}{l_A l_B} \sum_{i,j} I_{\{i=j\}} a_i b_j \quad (9)$$

This measure captures tree similarity but is not widely used, instead a measure of correlation  $r^2$  is typically reported. The measure  $M_{AB}$  is closely related to  $r^2$  since  $M_{AB}$  is exactly the probability of observing two markers (corresponding to trees  $A$  and  $B$ ) with  $r^2 = 1$ .

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, 36 and compare the correlation coefficient  $r^2$  of the resulting trees. Our results are estimates of the expected value of  $r^2$  based on  $x$  simulations of the process. Plots of the mean values of  $r^2$  are displayed in figure 7. Each of the plots correspond to simulations run from different sample sizes. In all cases the simulations are run until both of the loci have found a MRCA and then the resulting coalescent trees are used to compute  $r^2$ , where  $r^2$  is defined by (10). The measure 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.  $p_{11}$  denotes the probability of seeing the wild type in both trees and  $p_1$  and  $q_1$  denotes the probability of observing the first and second wildtype respectively.

$$r^2 = \frac{(p_{11} - p_1 q_1)^2}{p_1(1 - p_1)q_1(1 - q_1)} \quad (10)$$

**Figure 7** The mean correlation of trees for loci as recombinational distance between them is increased. Each plot corresponds to the simulation of different sample size (as labelled).

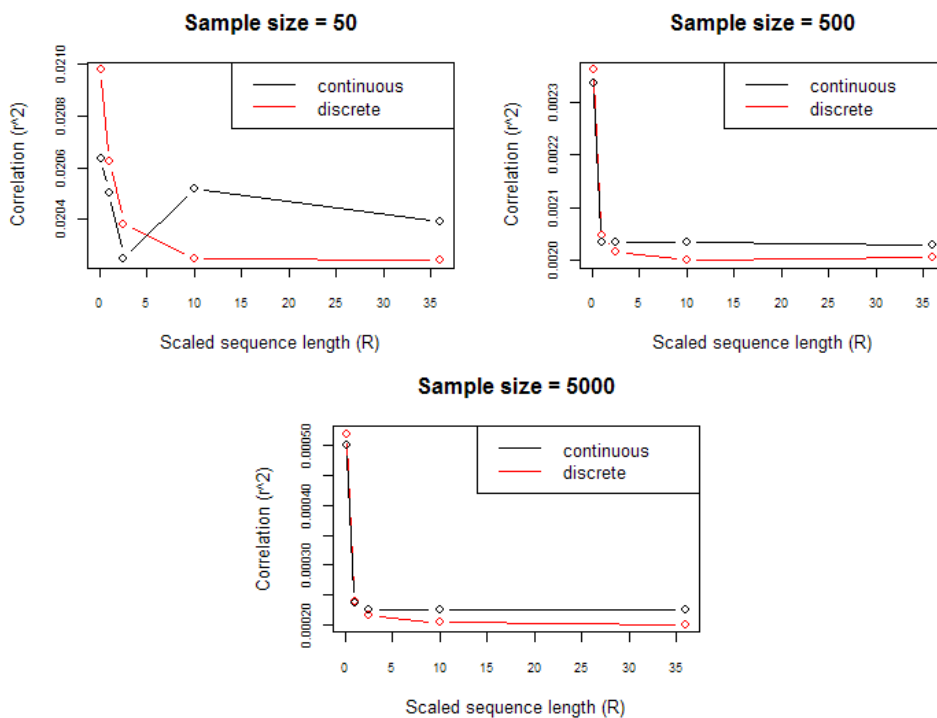


Figure 7 shows the value of  $r^2$  to five decimal places although visually it is possible to distinguish between the continuous and discrete models. In practice the values of  $r^2$  are the same to three or four decimal places and hence the differences displayed in figure 7 are negligible. We would expect the behaviour of  $r^2$  between two trees tracking the ancestral history of loci small recombinational distances apart to be the same. Locally the discrete and continuous processes agree. Differences (if any) are expected for larger values of  $R$ , however the values of  $r^2$  do not appear to be affected. The intuition for this can be gained from the plots in figure 4. 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 which place additional segments of ancestral material back onto an ancestor do so onto a region far away from an existing segment. It follows that the value of  $r^2$  decays in a similar way for both models.

### 3.6 The Impact of Gene Conversion

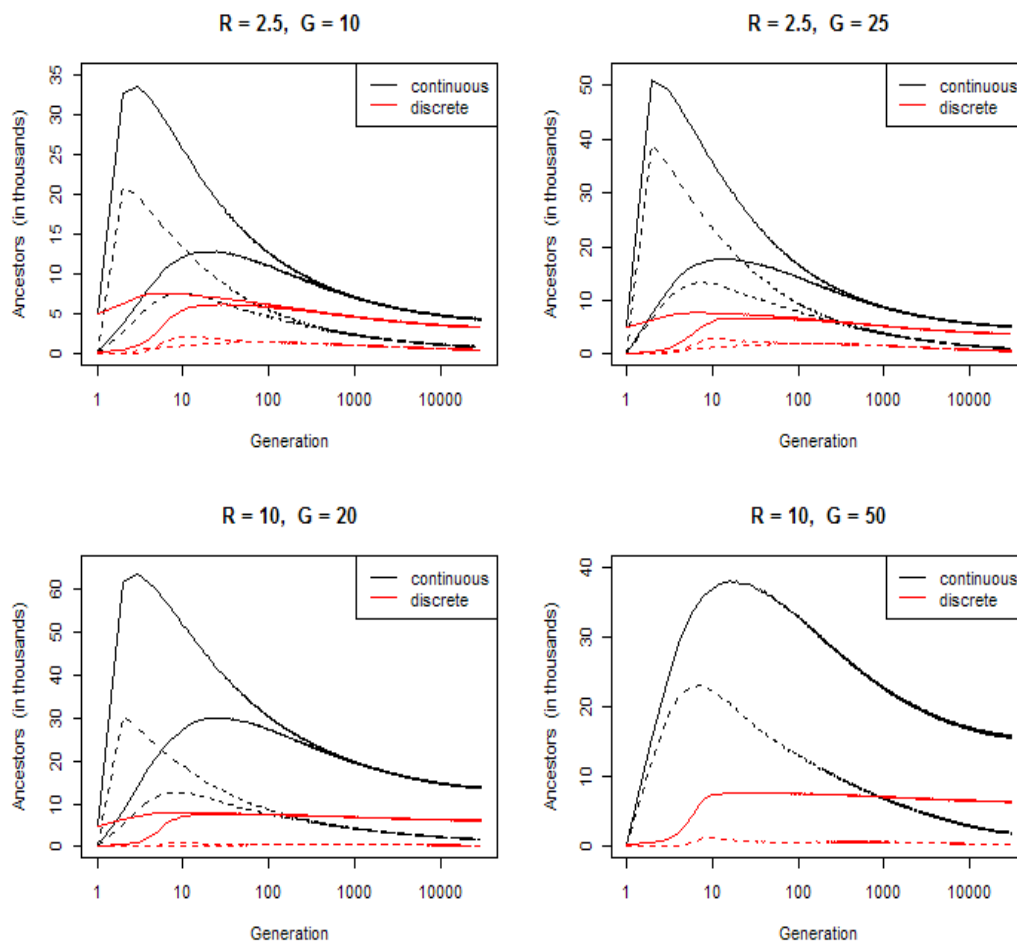
We compare the quantities discussed previously when gene conversion events are also simulated. We vary the ratio of rates, (gene conversion to crossover recombination) and use simulation to attain approximate results.

#### 3.6.1 The number of ancestors to a sample

The number of ancestors to an extant sample is a stochastic process as a function of time which converges to an equilibrium distribution. In section 3.1 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 8).

Each gene conversion event places two breakpoints (where possible) on a sequence, whereas a single crossover recombination event places exactly one break point 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 increasing (or adding) gene conversion has the effect of raising the over rate of recombination and our results in figure 5 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 which contain ancestral material only as a result of a gene conversion event (out of the total number of ancestors) rise.

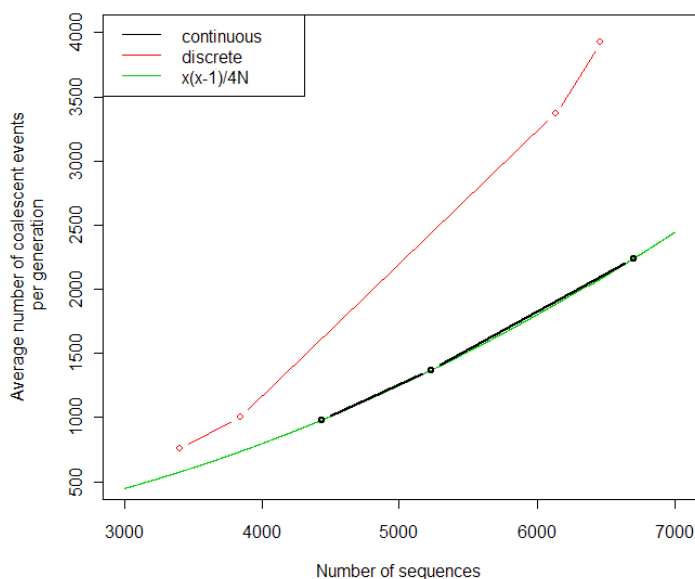
**Figure 8** The number of ancestors as a function of generations back in time. Each graph corresponds varying rates of recombination ( $R$ ) and gene conversion ( $G$ ) as labelled. The continuous model is plotted in black and the discrete model in red. The additional dotted lines represent the number of ancestors which only contain ancestral material as a result of a gene conversion event.



### 3.6.2 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 number of ancestors (in equilibrium). The rate of coalescence given the number of sequences remains the same for the discrete and continuous model. Our results in figure 9 are identical to those in figure 3 although the rates of gene conversion and crossover recombination used to simulate the results are different.

**Figure 9** The number of coalescent events per generation as a function of the number of sequences in the current generation. The results are generated using various different rates of gene conversion and crossover. Black lines correspond to the continuous model, red lines that of the discrete model and the green line the analytical expected number of events under the continuous model.

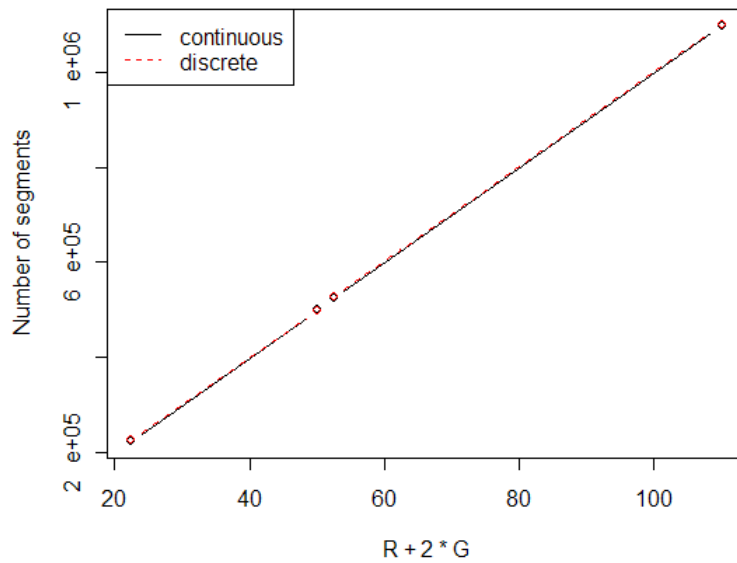


### 3.6.3 The distribution of ancestral material

The length of a segments 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 *onemillionth* 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. Other segments distributed by crossover events are typically longer in length and the average length of such segments is discussed in section 3.3 where only crossover events are simulated. The main difference we see 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 as in section 3.3. Each crossover event creates one break point and one extra segment, where as each gene conversion event creates two break points and two extra segments. For consistency with our results in figure 5 we would expect the average total number of segments to be  $R + 2G$  (since a conversion event generates twice as many new segments as a crossover event). Figure 10 confirms this is true and plotting the number of segments as a function of  $(R + 2G)$  yields the same straight line. Analogous to the results without gene conversion, the number of segments is not affected by type of model used. This is also as expected since the under estimation 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.

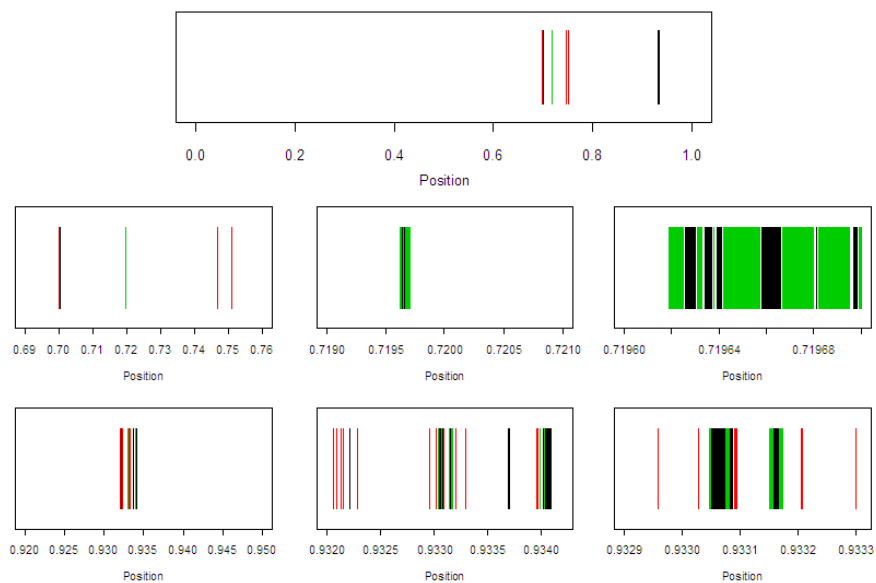
**Figure 10** The total number of ancestral segments as a function of  $R + 2G$ . The black line shows results for the continuous model and the red line that of the discrete model.



### 3.6.4 Ancestral Material on a Typical Common Ancestor

Investigation of the distribution of ancestral material in the absence of gene conversion is similar for the discrete and continuous models. On average more segments are found on a typical ancestor from the discrete model but the patterns displayed by any blocks containing ancestral are the same. With the addition of gene conversions this remains true although we see small gene conversion islands removed from what would have been a continuous segments formed by a crossover recombination events. This is illustrated in figure 11 by a typical ancestor selected at random with  $R = 2.5$  and  $G = 10$ .

**Figure 11** The distribution of ancestral material on a typical ancestor when the coalescent process is subject to cross over recombination events at rate 2.5 and gene conversion events at rate 10.



Investigation into the distribution of ancestral material in the presence of gene conversion raises further interesting

questions regarding the ancestry of each of the segments of ancestral material. Are the segments created for example by a gene conversion event or are they the result of a crossover event and a gene conversion event. Figure 8 shows the number of ancestors which contain ancestral segments which are created entirely from gene conversion events. Further specification of how ancestral segments are created may reveal further structures in the process. It is possible to label a segment of material with the events it has experienced. Although there are many combinations of possible events, further work may establish that there are a few distinct classes of events which occur frequently and many which rarely occur.

## 4 Conclusion

Our Monte Carlo simulations investigate the effect of using the continuous time approximation to the discrete coalescent when sample sequences are subject to recombination events. The quantities we calculate from our simulations can be considered in two classes; global or local. Global quantities are calculated across the whole sample (or current ancestral sample) and include the total number of ancestral sequences and the total rate at which coalescent events occur. Local quantities are calculated on a sequence level and include the number of segments of ancestral material, segment length and the patterns the segments of ancestral material form on a typical ancestor. The total amount of material ancestral to the sample is the sum of the length of the ancestral segments and although it can be considered a global quantity in the sense that it is a property of the whole sample, it can be decomposed as a sum of many local quantities.

Our simulation results show that the global quantities are affected by the neglect of possible simultaneous events in the continuous time model, while the local quantities are not. The total number of sequences ancestral to the sample is over estimated 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 are indistinguishable between models. The total number and 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 discrete model, but as explained above it can be considered as the sum of many local quantities.

Investigation into how the ancestral material is distributed on to the ancestors to the sample raises some interesting questions which require further work. Although the simulated patterns are not different between the discrete and continuous model (although there are on average more regions containing ancestral material on discrete model ancestor) it would be interesting to further classify the segments of ancestral material found on an ancestor according to the sequence of crossover/gene conversion recombination events and coalescent events which it experienced and investigate the distribution of the classes.

## References

- [1] R.Hudson, 1983, *Properties of the neutral allele model with intragenic recombination* : Theoretical Population Biology 23:183-201.
- [2] J. Hein, M.H. Schierup, C. Wiuf, 2005, *Gene Genealogies, Variation and Evolution*: Oxford University Press.
- [3] C. Wiuf, J. Hein, 1997, *On the number of ancestors to a DNA sequence* : Genetics 147: 1459-1468.
- [4] C. Wiuf, J. Hein, 1999, *Recombination as a point process along sequences* : Theoretical Population Biology 55(3), 248-259.
- [5] C. Wiuf, J.Hein, 1999, *The ancestry of a sample of sequences subject to recombination* : Genetics 151(3), 1217-1228.
- [6] J. F. C. Kingman, 1982, *The coalescent* : Stochastic Processes Applied. 13, 235-248.
- [7] J. F. C. Kingman, 1982, *On the genealogy of large populations* : Journal of Applied Probability 19A, 27-43.
- [8] Y. X. Fu, 2006, *Exact coalescence for the Wright-Fisher model*: Theoretical Population Biology (69) 385-394.
- [9] R.C. Griffiths, P. Marjoram, 1996, *Ancestral inference from samples of DNA sequences with recombination*: Journal of Computational Biology 3(4), 479-502.
- [10] J. Pitman, 1999, *Coalescents with multiple collisions* : Ann. Probability 27, 1870-1902
- [11] J. Schweinsberg, 2000, *Coalescents with simultaneous and multiple collisions* : Electronic Journal for Probability. (5) 1-50

- [12] S. Sagitov, 1999, *Coalescents with simultaneous multiple collisions* : Electronic Journal for Probability. (36) 1116-1125.
- [13] S. Sagitov, 2003, *Convergence to the coalescent with simultaneous multiple mergers* : Journal of Applied Probability. (40) 839-854.
- [14] N.O.T Sure ????, *Convergence to an equilibrium distribution* : where ever
- [15] C. Neuhauser, S.M. Krone 1997 *The genealogy of samples in models with selection* : Genetics 145(2), 519-534.