Theoretical Population Biology **57**, 357–367 (2000)
doi:10.1006/tpbi.2000.1462, available online at http://www.idealibrary.com on IDEAL®

A Coalescence Approach to Gene Conversion

Carsten Wiuf

Department of Statistics, University of Oxford, Oxford OX1 3TG, England

Received August 20, 1999

In this paper we develop a coalescent model with intralocus gene conversion. Such models are of increasing importance in the analysis of intralocus variability and linkage disequilibrium. We derive the distribution of the waiting time until a gene conversion event occurs in a sample in terms of the distribution of the length of the transferred segment, ζ. We do not assume any specific form of the distribution of ζ. Further, given that a gene conversion event occurs we find the distribution of (σ, τ), the end points of the transferred segment and derive results on correlations between local trees in positions χ_1 and χ_2. Among other results we show that the correlation between the branch lengths of two local trees in the coalescent with gene conversion (and no recombination) decreases toward a nonzero constant when the distance between χ_1 and χ_2 increases. Finally, we show that a model including both recombination and gene conversion might account for the lack of intralocus associations found in, e.g., *Drosophila melanogaster*.

Key Words: **coalescent model; gene conversion; intralocus variability; linkage disequilibrium; recombination.**

INTRODUCTION

This paper is concerned with modelling intralocus gene conversion within a coalescent framework. By gene conversion we refer to a genetic exchange where a short tract of information is transferred from one gamete to another without concurrent crossing-over (see, e.g., Andolfatto and Nordborg (1997) and references therein).

In the analysis of intralocus variability and linkage disequilibrium, the mechanisms operating at the sequence level must be properly understood. Recombination was first incorporated into the coalescent model (Kingman, 1982, describing the genealogical process of a sample of sequences taken from a population) by Hudson (1983) and has subsequently been investigated by a number of authors (Hudson and Kaplan, 1985, Griffiths and Marjoram, 1997, and Wiuf and Hein, 1997, among others).

The basis for Hudson's coalescent with recombination is the following. Over small intervals it may be assumed that crossing-over events are equally likely to occur at any point between two markers, and the probability of more than one event can be neglected. The probability of a crossing-over therefore increases linearly over small distances.

However, in models of homologous recombination the resolution of the Holliday junctions results either in a gene conversion with accompanying exchange of flanking regions (gene conversion with recombination) or in a gene conversion alone (Stahl, 1994). In the latter case a short tract of genetic information is transferred from one gamete to another without concurrent crossing-over. We refer to this phenomenon as gene conversion and the former phenomenon as recombination. Over small distances the effect of gene conversion events adds to the overall probability of producing a recombinant, an effect that is not accounted for in the coalescent with recombination alone.

Recently, it has been suggested that the lack of intralocus associations in regions of low rates of recombination found in, e.g., *Drosophila melanogaster* (Begun and Aquadro, 1995) is due to gene conversion (Andolfatto and Nordborg, 1997). Thus, there is a need to model gene conversion and study to what extent models with gene conversion and recombination can

0040-5809/00 $35.00

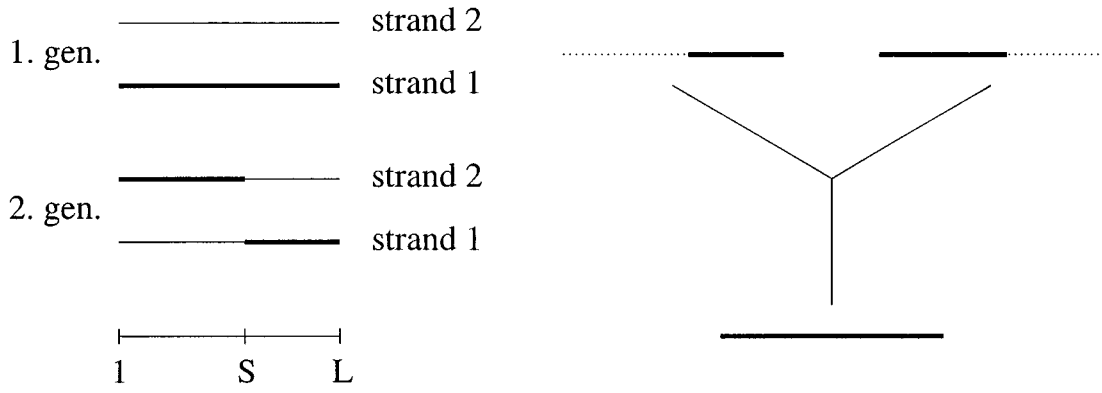

FIG. 1. Recombination. If the resolution of the Holliday junction results in an exchange of flanking regions we have (what here is called) a recombination event. In the left part of the figure, two of the four strands involved in the Holliday junction are shown. In the right part of the figure time starts at the present (the second generation) and goes backwards. Starting with either of the two strands/sequences in the second generation, the effect of the recombination event is to create two ancestors to the sequence; the positions to the left of position S share one ancestor and the positions to the right of S share another ancestor.

explain observed patterns of variability that recombination models apparently fail to explain.

In this paper we develop a gene conversion model within the coalescent framework. The model is very general in that it allows the distribution of the tract length to take an arbitrary form. The results are discussed relatively to the question raised by Andolfatto and Nordborg (1997).

The effect of recombination in the coalescent model is to break the material ancestral to a sequence up into two parts and distribute the parts onto two different ancestors, one carrying the ancestral material to the left of the recombination break point, S, the other carrying the material to the right of S (Fig. 1). In contrast, gene conversion as defined here breaks the material ancestral to a sequence at two points, S and T, and distribute the material to the left of S and that to the right of T onto one ancestor, the part in between S and T onto another ancestor (Fig. 2).

From the point of view of the ancestral graph (the graph describing the history of the sequences), the effect of a gene conversion event can be obtained by two recombination events and one coalescent event (Fig. 3).

Similarly, the effect of a recombination event can be obtained by one gene conversion event, where one end point falls outside the observed sequence.

More differences arise when we take a probabilistic view of the two processes. The probability of a gene conversion event might be very different from that of two recombination events followed by a coalescent event. Further, the latter (given that the first event occurs) will depend strongly on the current sample size; the higher

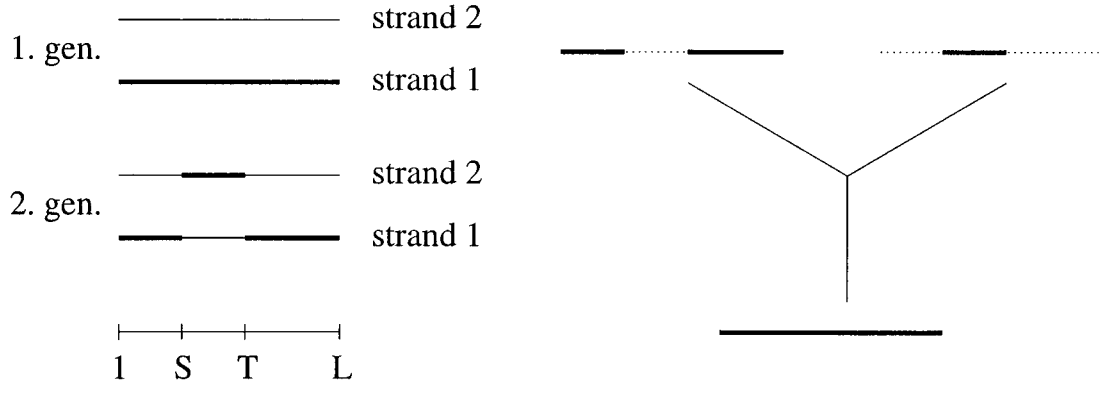

FIG. 2. Gene conversion. If the Holliday junction is resolved without an exchange of flanking regions we have a gene conversion event (without recombination). In the left part of the figure, two of the four strands involved in the Holliday junction is shown. In the right part of the figure time starts at the present (the second generation) and goes backwards. Starting with either of the two strands/sequences in the second generation, the effect of the gene conversion event is to create two ancestors to the sequence; the positions to the right of T and to the left of S share one ancestor and the position in between S and T share another ancestor.

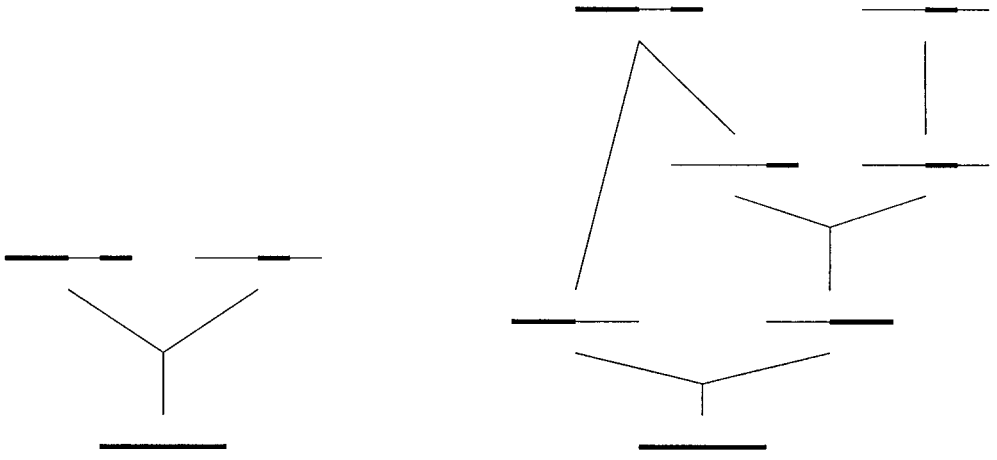

FIG. 3. Gene conversion versus recombination. The topological effect of one gene conversion event (shown to the left) can be obtained in a model with recombination only by two recombination events accompanied by a coalescence event. Assuming that the rate of gene conversion events is the same as the rate of recombination events, we find that the chance of two recombination events followed by a coalescence event is far smaller than that of a gene conversion event. The former also strongly depends on the number ancestral sequences present. Here, for example, given that a coalescence event occurs, only one out of three possible mergings will result in the desired distribution of ancestral material.

the sample size, the lower the chance that the two recombined sequences will coalesce before coalescing with any other sequence in the sample.

As sequence length increases, gene conversion events where only one end point falls within the observed sequence length become rare, in contrast to recombination models where this is the only type of event. Further, for an event of this type the gene conversion end point will be distributed just around the end points of the sequence, again in contrast to recombination models where the break point is uniform on sequence length (standard assumption; see, e.g., Griffiths and Marjoram, 1997).

The distribution of the gene conversion end points, S and T, is determined by the distribution of the length of the transferred chunk between S and T. Thus, we should not expect S and T to be uniformly distributed, even in the case where only one end point is within the observed sequence length.

In the eighties, models of gene conversion within multigene families were developed (see, e.g., Ohta, 1986). These models do not apply to intralocus gene conversion, and the model scheme differs in several ways from the one proposed here. Hudson (1994) discusses models with various genetic transformations similar to gene conversion. In these models blocks of genes are transferred and, further, the model scheme is different in a number of places from the one proposed here.

A specific example of the model developed here was investigated in a previous paper (Wiuf and Hein, 2000) and we refer to that paper for further discussion and implications of gene conversion.

THE MODEL

The model within the coalescent framework of a population subject to gene conversion is the following. The population is constantly of size N and diploid; i.e., the effective population size is $2N$. A new generation is obtained from the present generation by sampling $2N$ sequences with replacement forming random pairs of sequences, and letting one of them transfer a short tract of nucleotides to the other sequence. The mode of transfer is described below.

Sequences are of length $L+1$ nucleotides, so there are L gaps between nucleotides. Consider one such sequence. Assume that in any generation the probability of gene conversion initiating between any two positions in the sequence is g, independent of whether gene conversions initiate elsewhere along the sequence. The probability distribution of the number of conversions is binomial with parameters g and L, i.e., $Bin(g, L)$, so that the probability of more than one gene conversion in one generation is negligible for small g. Put $G = 4gNL$; G is defined similarly to the rate of recombination in the coalescent model with recombination. If N is large and gL small, the number of sequences undergoing a gene conversion in one generation is Poisson with intensity $G/2$. Where along the sequence a gene conversion initiates is uniformly distributed among all sites.

The transferred chunk of nucleotides originates from a randomly chosen sequence in the population. Let the length, Z, of the converted chunk have distribution $P(Z = i \mid \text{conversion})$, $i \geqslant 1$; i.e., at least one nucleotide is transferred. It is assumed that the insertion happens to

the right of the gap where the conversion initiates. If the conversion initiates at gap S, $S = 1, 2, ..., L$ (S for start), then the end point of the conversion is $T = S + Z$ (T for terminate). If $S + Z$ is greater than L, the conversion falls partly outside the $L + 1$ observed nucleotides.

Alternatively, we could consider a model where the insertion happens to the right of the gap with probability p, $0 \leqslant p \leqslant 1$, and to the left with probability $1 - p$, independent of the position where it happens. In this case the gene conversion will be partly outside the $L + 1$ observed nucleotides if $S - Z$ is smaller than 1 or if $S + Z$ is greater than L. It turns out that this model can be considered equivalent to that mentioned above (corresponds to $p = 1$), and as we go along, we will make this point clear.

If g is not sufficiently small so that more than one conversion can happen in one sequence in one generation, overlapping gene conversions could result. However, our interest is small values of g and the diffusion limit considered ensures that only one conversion event can happen at a time. For larger values of g, one can circumvent the problem by truncating the variable Z, so that no overlaps occur.

The genealogical process of a sample of n present-day sequences will be studied; time will start at the present and increase going backwards in time. Under the above assumptions we find that the waiting time, W_C, measured in units of $2N$ generations until a sequence has been created by a gene conversion event that initiates within the sequence is exponentially distributed with parameter $G/2$; i.e., $W_C \sim \text{Exp}(G/2)$, if N is large and gL small. The rate of gene conversions initiating outside the sequence but ending within the observed sequence must also be taken into account. This rate turns out to depend on the distribution of Z. The rate of coalescence is $n(n-1)/2$ if there are n sequences in a sample (Kingman, 1982).

We will now explore the structure of the gene conversion in more detail. Consider a sequence in a certain generation. Let C denote the event that a gene conversion happens in a given sequence and in a given generation. Further, let C_1 denote the event that the gene conversion falls partly outside the $L + 1$ observed nucleotides, that is, $S + Z > L$, and let C_2 denote the event that both end points are within the sequence length (lower index i indicates that i, $i = 1, 2$, of the two end points of the converted chunk are within the L nucleotides).

Denote by ζ, σ, and τ, respectively, the variables Z/L, S/L, and T/L, respectively. Though σ and τ depend on L, e.g., they both take values in $\{1/L, 2, ..., 1 - 1/L\}$, we suppress L in the notation. The probability measure wrt to the model with sequences of length $L + 1$ is denoted P_L. In terms of ζ, σ, and τ, the sequences are considered as part of the interval $(0, 1)$, and as L increases the nucleotides become more densely spaced in $(0, 1)$.

Assume that as L tends to infinity the distribution of $\zeta = Z/L$, $P_L(\zeta \leqslant \cdot \mid C)$, converges weakly;

$$P_L(\zeta \leqslant z \mid C) \to P(\zeta \leqslant z \mid C) \tag{1}$$

for all points $z > 0$ where $P(\zeta \leqslant \cdot \mid C)$ is continuous. In the following, whenever the limit as $L \to \infty$ is considered, we assume that z is a continuity point of $P(\zeta \leqslant \cdot \mid C)$. All results will be given in terms of $P(\zeta \leqslant \cdot \mid C)$ and $P(\zeta > \cdot \mid C)$, which both are right continuous, and the limit expressions are valid for all $0 \leqslant z < 1$ irrespective of whether z is a continuity point or not.

A few examples are appropriate:

(1) All gene conversions have the same length, $Z = Z_0$; $P_L(Z \leqslant i \mid C) = 1_{[Z_0, \infty)}(i)$, where $1_A(\cdot)$ is the indicator function of a set, and $P_L(\zeta \leqslant z \mid C) \to 1_{[\zeta_0, \infty)}(z)$, $z \neq \zeta_0$, if $Z_0/L \to \zeta_0$ as $L \to \infty$. The limit distribution of ζ is the Dirac measure at ζ_0, the measure with all mass in ζ_0.

(2) Z follows a negative binomial distribution with parameters $1 > q > 0$ and $\kappa > 0$;

$$P_L(Z = i \mid C) = \binom{i + \kappa - 1}{i} (1 - q)^i q^\kappa, \qquad i \geqslant 0$$

and

$$P_L(\zeta \leqslant z \mid C) \to \int_0^z \frac{Q^\kappa}{\Gamma(\kappa)} x^{\kappa - 1} \exp(-Qx)\, dx, \qquad z > 0,$$

if $qL \to Q$ for $L \to \infty$. The limit distribution of ζ is $\Gamma(\kappa, Q)$. This example includes the geometric distribution ($\kappa = 1$) as a special case with exponential limit, $\Gamma(1, Q) = \text{Exp}(Q)$. Hilliker *et al.* (1994) find that the geometric distribution fits well to *Drosophila melanogaster* data and Betrán *et al.* (1997) support this conclusion in *Drosophila subobscura*.

The two examples represent two opposite cases, one in which the variance of ζ is zero (Example 1) whereas in Example 2 the variance of ζ is κ/Q^2 and thus can be arbitrary large. The ratio of the mean to the variance of the negative binomial variable, Z, is always larger than one, whereas for the limit variable, ζ, the ratio can take all values. This is due to the fact that ζ is Z divided by L which can take large values.

We find

$$P_L(C_2 \mid C) = \frac{1}{L}\sum_{i=1}^{L} P_L(Z \leqslant L-i \mid C) \to \int_0^1 P(\zeta \leqslant x \mid C)\, dx = 1 - E_C[\zeta \wedge 1], \qquad (2)$$

using (1) and where $E_A[\,\cdot\,]$ denotes expectation wrt P given the event A. The variable $x \wedge y$ denotes the minimum of x and y.

From (2), $C_1 \cup C_2 = C$, and $C_1 \cap C_2 = \varnothing$ we have

$$P_L(C_1 \mid C) = 1 - P_L(C_2 \mid C) \to E_C[\zeta \wedge 1].$$

Denote by W_{C_i} the time until one sequence is created by an event of type C_i, $i = 1, 2$. We have $W_{C_1} \sim \mathrm{Exp}(GE_C[\zeta \wedge 1]/2)$ and $W_{C_2} \sim \mathrm{Exp}(G(1 - E_C[\zeta \wedge 1])/2)$, such that $W_C = \min(W_{C_1}, W_{C_2}) \sim \mathrm{Exp}(G/2)$.

For an event in C_2, the distributions of σ, τ, and (σ, τ) and that of $\zeta = \sigma - \tau$ are of importance, and similarly for an event in C_1, the distribution of σ is of importance (the end point τ of the gene conversion falls outside the sequence considered).

The distribution of σ, conditional on C_2, can be found as follows.

$$L \cdot P_L(\sigma = s \mid C_2) = \frac{L \cdot P_L(\sigma = s, C_2 \mid C)}{P_L(C_2 \mid C)} = \frac{P_L(Z \leqslant L - sL \mid C)}{P_L(C_2 \mid C)} \to \frac{P(\zeta \leqslant 1 - s \mid C)}{1 - E_C[\zeta \wedge 1]} = f_\sigma(s \mid C_2), \qquad (3)$$

$0 < s < 1$, because $S + Z \leqslant L$ on C_2. The function $f_\sigma(s \mid C_2)$ is the density of σ wrt the Lebesgue measure on $(0, 1)$.

Similarly, we find

$$L \cdot P_L(\tau = t \mid C_2) = \frac{P_L(Z \leqslant tL - 1 \mid C)}{P_L(C_2 \mid C)} \to \frac{P(\zeta \leqslant t \mid C)}{1 - E_C[\zeta \wedge 1]} = f_\tau(t \mid C_2), \qquad (4)$$

$0 < t < 1$, and $f_\tau(t \mid C_2)$ is the density of τ wrt the Lebesgue measure on $(0, 1)$.

We note that $f_\sigma(s \mid C_2) = f_\tau(1 - s \mid C_2)$ and hence $\sigma \sim 1 - \tau$ (where $\sim$ denotes "is distributed like"). The reason σ and τ both are continuous variables is essentially that the initiation point of a gene conversion is uniform along $(0, 1)$.

Next, consider the distribution of $\zeta = \tau - \sigma$. In general ζ conditional on C_2 will not be continuous (cf. Example 2). Proceeding as above, we find

$$P_L(\zeta \leqslant z \mid C_2) = \frac{P_L(\zeta \leqslant z, C_2 \mid C)}{P_L(C_2 \mid C)} = \frac{1}{L}\sum_{j=1}^{zL}\sum_{i=1}^{L-j} \frac{P_L(Z = j \mid C)}{P_L(C_2 \mid C)} \to \left\{1 - \int_0^z P(\zeta > x \mid C)\, dx - (1 - z)\, P(\zeta > z \mid C)\right\} \times (1 - E_C[\zeta \wedge 1])^{-1}, \qquad (5)$$

$0 < z < 1$, by interchanging the order of summation and performing calculations similar to those of Eq. (2).

If the limit distribution of ζ, conditioned on C_2, has density $f_\zeta(z \mid C_2)$ wrt the Lebesgue measure it can be found by differentiation of (5). In that case we find a simpler expression for (5),

$$f_\zeta(z \mid C_2) = \frac{(1 - z)\, f_\zeta(z \mid C)}{1 - E_C[\zeta \wedge 1]},$$

with $0 < z < 1$. However, as Example 1 showed, there are cases of interest that do not allow such a simplification.

Finally, we derive the joint distribution of (σ, τ). The distribution relates to (5) and (6) because $\zeta = \tau - \sigma$, and therefore (σ, τ) might not have a density wrt the Lebesgue measure on $(0, 1)^2$. In general we have with $0 < s \leqslant t < 1$

$$P_L(\sigma \leqslant s, \tau \leqslant t \mid C_2) = \sum_{j=1}^{sL}\sum_{i=j+1}^{tL} \frac{P_L(S = j, T = i \mid C)}{P_L(C_2 \mid C)} \to \left\{s - \int_0^s P(\zeta > t - x \mid C)\, dx\right\} (1 - E_C[\zeta \wedge 1])^{-1}. \qquad (7)$$

The calculations are similar to those of Eq. (2). This distribution is invariant under reflection in $x = 1/2$; i.e., the distribution of $(1 - \tau, 1 - \sigma)$ is also given by (7). This is

easily seen in the special case where (σ, τ), conditioned on C_2, has density

$$f_{\sigma,\tau}(s, t \mid C_2) = \frac{f_\zeta(t-s \mid C)}{1 - E_C[\zeta \wedge 1]} \tag{8}$$

by differentiation of (7) wrt s and t. Otherwise, the distribution of $(1-\tau, 1-\sigma)$ can be found from Eq. (7), using that the right hand side of (7) is continuous in both s and t.

Consider the alternative model, where a gene conversion could happen to the right of the initiation point with probability p and to the left of the initiation point with probability $1-p$. If we let s and t, $s<t$, denote the end points of the inserted chunk, we will find the same distributions as above. In this case, s will then be the starting point with probability p and the termination point with probability $1-p$. Similarly for the point t.

Let us now turn to the distribution of σ conditional on C_1, that only the starting point is within the L nucleotides. Since the distribution of σ conditional on C is uniform on $(0, 1)$ we have $1 = f_\sigma(s \mid C_2)\, P(C_2 \mid C) + f_\sigma(s \mid C_1)\, P(C_1 \mid C)$ or

$$f_\sigma(s \mid C_1) = \frac{P(\zeta > 1-s \mid C)}{E_C[\zeta \wedge 1]} \tag{9}$$

for $0<s<1$, and where $f_\sigma(s \mid C_1)$ is the density of σ wrt the Lebesgue measure on $(0, 1)$.

In the alternative model, the distribution of s would be $pf_\sigma(s \mid C_1) + (1-p)\, f_\sigma(1-s \mid C_1)$, because with probability p the conversion would go to the right and with probability $1-p$ to the left. However, this difference between the two models vanishes due to the below.

Gene conversions initiating outside the $L+1$ nucleotides will have a chance of terminating within the $L+1$ observed nucleotides. Assume that the entire chromosome potentially consists of an infinite array of sequences of length L plus the observed one of length $L+1$, and that the gene conversion model described above is valid for the entire chromosome.

Number the sequences to the left of the observed $L+1$ nucleotides, $S_1, S_2, \ldots$, such that S_n is the nth sequence to the left of the observed $L+1$ nucleotides, here called S_0. Besides the L nucleotides, we let S_n, $n=1, 2, \ldots$, consist of the $L-1$ gaps between the L nucleotides and the gap just to the right of nucleotide L, i.e., the gap between sequences S_n and S_{n-1}, $n=1, 2, \ldots$, in total L gaps. First note that a gene conversion initiating in S_{n+1}, $n=0, 1, \ldots$, has probability

$$Q_{L,n} = \frac{1}{L} \sum_{i=1}^{L} P_L(nL+i \leqslant Z < (n+1)\, L+i \mid C) \tag{10}$$

of terminating within S_0. The waiting time until a gene conversion event initiating within S_{n+1} and ending within S_0 is approximately exponential with intensity $Q_{L,n}G/2$, so that the process is well defined iff $\lim_L \sum_n Q_{L,n} < \infty$, i.e., the gene conversions affecting S_0 do not arrive instantaneously, but are spread out in time. Let C_o (o for outside) denote the event that a gene conversion initiates outside S_0 and terminates within S_0. Clearly,

$$P_L(C_o) = \sum_{n=0}^{\infty} Q_{L,n} = \frac{1}{L} \sum_{i=1}^{L} P_L(Z \geqslant i \mid C) \to E_C[\zeta \wedge 1], \tag{11}$$

where we have used (10). We find that both $P_L(C_o)$ and $P_L(C_1 \mid C)$ converge to $E_C[\zeta \wedge 1]$, and that the waiting time, W_{C_o}, until a gene conversion initiating outside S_0 ends within S_0 is properly defined. The distribution of W_{C_o} is exponential with intensity $GE_C[\zeta \wedge 1]/2$, that is, distributed like the waiting time, W_{C_1}, until an event of type C_1; $W_{C_o} \sim W_{C_1}$.

Similarly to the previous proofs, we find the distribution of the terminating point, τ, conditional on C_o,

$$L \cdot P_L(\tau = t \mid C_o) \to \frac{P(\zeta > t \mid C)}{E_C[\zeta \wedge 1]} = f_\tau(t \mid C_o), \tag{12}$$

$0<t<1$, and $f_\tau(t \mid C_o)$ is the density of τ wrt the Lebesgue measure on $(0, 1)$.

The events C_o and C_1 are independent. Thus, the waiting time $W_{C_1 \cup C_o} = \min(W_{C_1}, W_{C_o})$ until an event of either type C_1 or type C_o is exponentially distributed with parameter $GE_C[\zeta \wedge 1]$, and where the initiation/termination point, σ, happens is distributed with density

$$L \cdot P(s \mid C_1 \cup C_o) \to \frac{1}{2} \{ f_\sigma(s \mid C_1) + f_\tau(s \mid C_o) \}$$

$$= \frac{P(\zeta > 1-s \mid C) + P(\zeta > s \mid C)}{2E_C[\zeta \wedge 1]}; \tag{13}$$

cf. Eqs. (9) and (12).

In the alternative model, we could derive the same result, and the two models discussed here are identical wrt the distributions of waiting times affecting the history of a sequence of length $L+1$, and where along the sequence breaks occur. Note that in the alternative model, we would also have to consider the infinite array of sequences to the right of S_0.

We recapitulate these results in the following theorem.

THEOREM. *The waiting time, W, until a sequence of length $L+1$ nucleotides is created by a gene conversion is exponentially distributed*

$$W \sim \text{Exp}\left(\frac{G}{2}\{1+E_C[\zeta \wedge 1]\}\right), \tag{14}$$

where $G=4NLg$, and g is the probability of a gene conversion initiating between any two nucleotides. The variable ζ is the normed length of the transferred nucleotide, Z; i.e., $\zeta = Z/L$. The waiting time W is the minimum of three independent waiting times: the time until a gene conversion event with both end points within the sequence, the time until an event that starts within the sequence and ends outside, and the time until an event that starts outside and terminates inside the sequence.

Given that a gene conversion occurs, the probability that both, end points of the inserted chunk are within the sequence is

$$p_2 = \frac{1-E_C[\zeta \wedge 1]}{1+E_C[\zeta \wedge 1]},$$

and the probability that only one end point is within the sequence is

$$p_1 = \frac{2E_C[\zeta \wedge 1]}{1+E_C[\zeta \wedge 1]}.$$

Given that the gene conversion is entirely within the sequence, the end points, σ and τ with $\sigma < \tau$, have distributions given by (3), (4), *and* (7).

Given that only one end point is within the sequence, this end point, σ, has distribution given by (13). *The end point can both be an initiation point as well as a termination point of the gene conversion.*

If the length of the sequence, i.e., the number of nucleotides, is multiplied by a factor, λ, G becomes multiplied by λ because $G=4gNL$. Consider the model with parameter $G'=\lambda G$ relatively to that of G. Quantities marked with a prime refer to the model with parameter $G'=4gNL'=\lambda G$, and unmarked quantities refer to the model with parameter G. We have $\zeta' = \zeta/\lambda$ and thus the waiting time until a sequence is created by a gene conversion event is distributed

$$W \sim \text{Exp}\left(\frac{\lambda G}{2}\left\{1+\frac{1}{\lambda}E_C[\zeta \wedge \lambda])\right\}\right)$$
$$\sim \text{Exp}\left(\frac{\lambda G}{2}+\frac{G}{2}E_C[\zeta \wedge \lambda])\right). \tag{15}$$

Assume $0 < E_C[\zeta] < \infty$. If $\lambda \to \infty$, then $E_C[\zeta \wedge \lambda]/\lambda \to 0$ and $\lambda G E_C[\zeta \wedge \lambda]/2\lambda \to G E_C[\zeta]$. Also $p_2 \to 1$. This is expected because the term $\lambda G E_C[\zeta \wedge \lambda]/2\lambda$ is the rate of gene conversion initiating outside the $L'+1$ nucleotides and as $L'=\lambda L$ increases this type of gene conversions will affect the two ends of the sequence only. Thus, this type becomes more rare compared to type C_2 events. Similarly, if $\lambda \to 0$, then $E_C[\zeta \wedge \lambda]/\lambda \to 1$, $\lambda G E_C[\zeta \wedge \lambda]/2\lambda \to 0$, and $p_2 \to 0$. So most events are of type $C_1 \cup C_o$. Again this is expected.

This parameterization has the advantage that we easily can study and compare effects of gene conversion in samples of different nucleotide lengths but with similar value of g. The parameter λ can be considered the length of the sequences as it increases linearly with the number of nucleotides in the sequences. It is convenient to let $G=1$ but $G=1/2$ is also advantageous. In the latter case sequence length is measured in expected number of gene conversions per sequence per $2N$ generations.

RESULTS

Throughout the section the results are stated in terms of λG so that easy comparisons between samples of different sequence length can be made.

Define $G_{n,1}$ as the number of gene conversions affecting the history of a sample of size n, and $G_{n,2}$ as the number of end points of gene conversions affecting the history of the sample. Clearly, $G_{n,1} \leqslant G_{n,2}$. Both variables are of interest; $G_{n,2}$ is the number of positions where the history of the sequences potentially changes. We write "potentially" because the next event following a gene conversion event might be a coalescence involving the two sequences created by the gene conversion event, thereby erasing the possibility that the gene conversion affects the history of the sequences. In the coalescent with recombination the two numbers $G_{n,1}$ and $G_{n,2}$ are identical.

The variable $G_{n,2}$ is an additive function in the number of nucleotides, whereas $G_{n,1}$ is not. Divide a sample of sequences into two samples of length $L/2$ and let $G_{n,1}(i)$, $i=1, 2$, denote the number $G_{n,1}$ in each half. If a gene conversion initiates in the first half and terminates in the second half, it will add to both $G_{n,1}(1)$ and $G_{n,1}(2)$, but only be counted once in $G_{n,1}$. Thus, $G_{n,1}$ is not additive. In contrast, both points will add to $G_{n,2}$ and in fact $G_{n,2}$ fulfills $G_{n,2} = G_{n,2}(1) + G_{n,2}(2)$.

The expected value of $G_{n,2}$ is

$$E[G_{n,2}] = 2\lambda G \sum_{i=1}^{n-1} \frac{1}{i}. \tag{16}$$

and a lower bound to the expected value of $G_{n,1}$ is

$$E[G_{n,1}] > G(\lambda + E_C[\zeta \wedge \lambda]) \sum_{i=1}^{n-1} \frac{1}{i}. \qquad (17)$$

These results can be obtained by considering the rate at which different events happens in a given position s. For example, the rate, r, by which gene conversions of type C_2 initiates in a small region of size ds around position s is $r = \frac{G}{2} P(\zeta \leqslant 1 - s \mid C)\, ds$ (see the Theorem and (3)). The expected number of such events within the region $s + ds$ is then given by $rE[B] = 2r \sum_i 1/i$, where B is the total branch length of the genealogy in position s. Finally, integration over s gives the expected number along the sequences. The bound for the expectation of $G_{n,1}$ is obtained by counting all events of type C_1 and C_o plus all type C_2 initiation points within ancestral material. Left out are type C_2 termination points within ancestral material for which the initiation point is outside the ancestral material.

Also of interest is the probability that no gene conversions occur before the most recent common ancestor (MRCA) of the sample. We find

$$P(G_{n,1} = 0) = P(G_{n,2} = 0) = \prod_{i=1}^{n-1} \frac{i}{i + G(\lambda + E_C[\zeta \wedge \lambda])}. \qquad (18)$$

Other quantities of interest can be found from similar results obtained for the coalescent with recombination. We will mention a few of these. Consider two positions, χ_1 and χ_2, in a sample of n sequences. Let the distance between the two positions be λ. We are interested in the trees, $T_n(\chi_1)$ and $T_n(\chi_2)$, that describe the sequences at χ_1 and χ_2. Only events of types C_1 and C_o in between positions χ_1 and χ_2 affect the relation between the two trees. Events of type C_2 in between the positions cannot be traced. The rate, $r_\lambda/2$, by which events of type C_1 and C_o happens is, per sequence (see the Theorem),

$$r_\lambda = 2\lambda G \frac{1}{\lambda} E_C[\zeta \wedge \lambda] = 2GE_C[\zeta \wedge \lambda]. \qquad (19)$$

Let $B_n(\chi)$ be the total branch length of $T_n(\chi)$. Applying results in Griffiths (1991) we find, for $n = 2$,

$$\mathrm{Cov}(B_2(\chi_1), B_2(\chi_2)) = \frac{4(18 + r_\lambda)}{18 + 13r_\lambda + r_\lambda^2}, \qquad (20)$$

where Cov denotes the covariance between variables. We note that the expression in (20) obtains its minimum when $\lambda = \infty$, in which case $r_\infty = 2GE_C[\zeta]$. Thus the covariance is strictly positive for all values of λ and converges towards a non-zero constant as $\lambda \to \infty$ (unless $E_C[\zeta] = \infty$, which is not realistic biologically).

Hudson (1983) gives an approximate formula for the covariance for a general n. He finds (with r_λ inserted)

$$\mathrm{Cov}(B_n(\chi_1), B_n(\chi_2)) \approx \frac{4(18 + r_\lambda)}{18 + 13r_\lambda + r_\lambda^2} \sum_{i=1}^{n-1} \frac{1}{i^2}. \qquad (21)$$

The approximation is exact for $n = 2$ and as believed to be fairly good for small and moderate values of r_λ for all n (Kaplan and Hudson, 1985).

Let $H_n(\chi)$ be the height of $T_n(\chi)$. A lower bound to $P(H_n(\chi_1) = H_n(\chi_2))$ can be found:

$$P(H_n(\chi_1) = H_n(\chi_2)) > P(G_{n,2} = 0) = \prod_{i=1}^{n-1} \frac{i}{i + G(\lambda + E_C[\zeta \wedge \lambda])}. \qquad (22)$$

For $n = 2$ we have $P(H_2(\chi_1) = H_2(\chi_2)) = \mathrm{Cov}(B_2(\chi_1), B_2(\chi_2))/4$ (Griffiths, 1991).

AN EXAMPLE

In most applications a gene conversion model will have limited use on its own. What is needed is a combination of the coalescent with gene conversion and that of recombination. In the following we discuss a few aspects of such a combined model and relate the results to data from *Drosophila melanogaster*.

Let the recombination rate be given by λR, where $R = 4rNL$, with r being the probability of a recombination between any two sites per generation, and let the corresponding gene conversion rate be λG, $G = 4gNL$. Consider two positions χ_1 and χ_2 at distance λ. Recombinants are produced at a rate $\rho_\lambda/2$, where

$$\rho_\lambda = r_\lambda + R\lambda = 2GE_C[\zeta \wedge \lambda] + R\lambda \qquad (23)$$

(see (19)). For small λ, $r_\lambda = 2G\lambda + o(\lambda)$, where $o(\lambda)$ denotes a function such that $o(\lambda)/\lambda \to 0$ for $\lambda \to 0$. We find

$$\rho_\lambda \approx (2G + R)\, \lambda. \qquad (24)$$

For large λ the gene conversion effect tends to disappear,

$$\rho_\lambda \approx r_\infty + R\lambda = 2GE_C[\zeta] + R\lambda, \qquad (25)$$

which is of order $R\lambda$.

Andolfatto and Nordborg (1997) (and references therein) give the following estimates of g and r in *Drosophila melanogaster*,

$$g \approx 3 \times 10^{-8} \quad \text{and} \quad r \approx 1 \times 10^{-8}, \qquad (26)$$

and Hilliker *et al.* (1994) provide an estimate of the expected tract length, $E_C[Z] \approx 10^{+3}/3$ nucleotides. Further, put $2N = 10^6$ and $G = 3$ so that $R = 1$ and length is measured in expected number of recombinations per sequence per $4N$ generations.

Equations (24) and (25) apply immediately but to find ρ_λ for all λ we must adopt a model of the tract length, Z. Here Example 1 is chosen for convenience: Z is constant, $Z = Z_0 = 350$ nucleotides. A different choice of distribution does not alter the results significantly. The rate ρ_λ becomes (see (23))

$$\rho_\lambda = 2(4gNZ_0 \wedge 6\lambda) + R\lambda = 40 \wedge G\lambda + \lambda \qquad (27)$$

because $G\zeta = 4gNLZ_0/L = 4gNZ_0 = 20$. The rate ρ_λ for small λ is considerably larger than the rate λ in a pure recombination model. When $\lambda = 40$ the contribution from recombination equals that from gene conversion and in that case the length of the sequence is $40/(4rN) = 2.000$ nucleotides. This might very well explain the lack of intralocus associations reported in

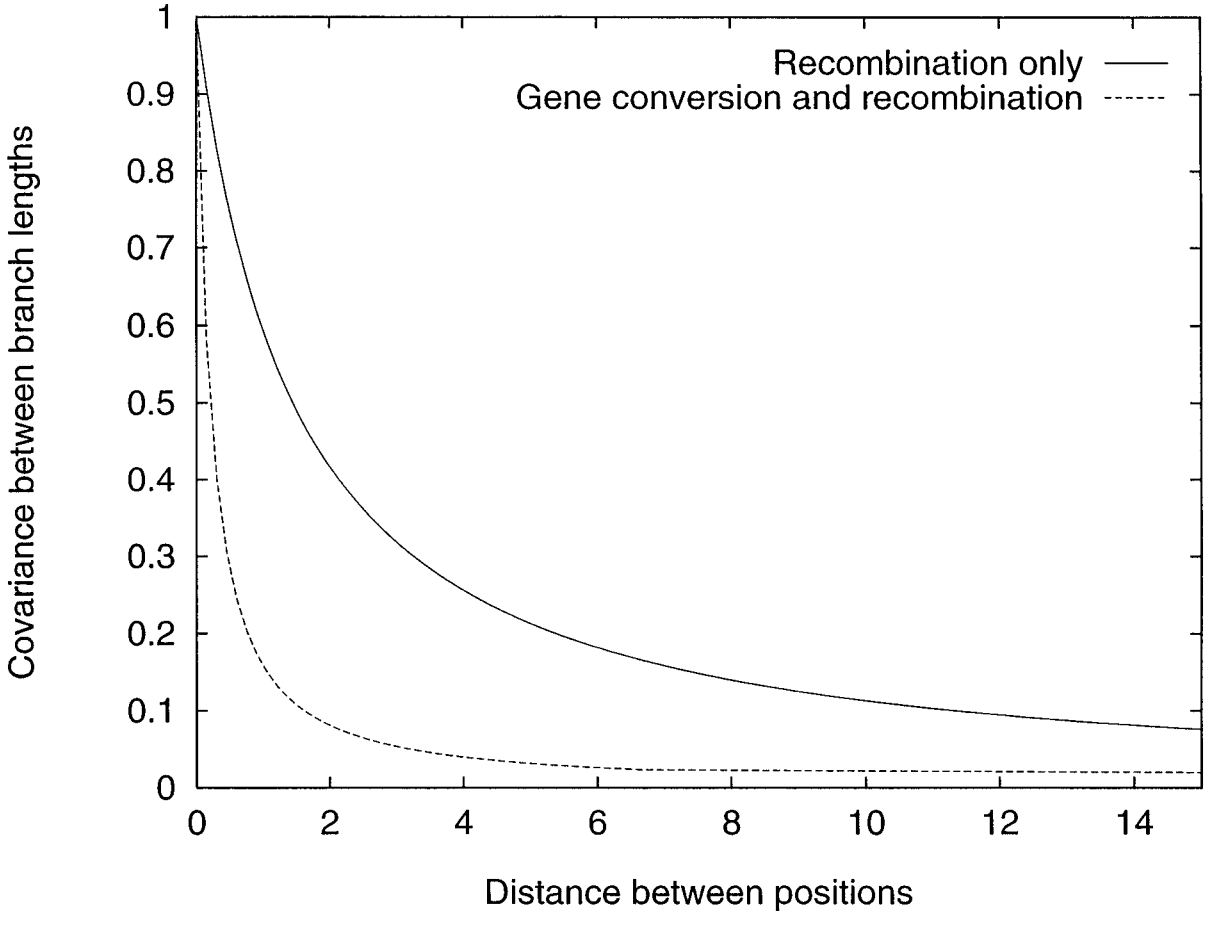

FIG. 4. Covariance between branch lengths. Sequence length is in expected number of recombination events per sequence per $4N$ generations. The figure shows the normed covariance, $\mathrm{Cov}(B_2(\chi_1), B_2(\chi_2))/4$, between the branch lengths, $B_2(\chi_i)$, $i = 1, 2$, in two positions, χ_1 and χ_2, at distance λ in a sample of size 2. According to Hudson (1983) this is an accurate approximation to the covariance for $n > 2$ up to a scaling factor (dependent on the sample size, n). Note that for $n = 2$, $\mathrm{Cov}(B_2(\chi_1), B_2(\chi_2))/4 = \mathrm{Cov}(H_2(\chi_1), H_2(\chi_2))$, where $H_2(\chi_i)$ is the height of the tree in position χ_i, $i = 1, 2$. The values of the rates of gene conversion and recombination are those estimated in *Drosophila melanogaster* (Andolfatto and Nordborg, 1997, and references therein). The estimated length of a gene conversion tract is 350 nucleotides (Andolfatto and Nordborg, 1997, and references therein). The distance $\lambda = 15$ corresponds to a sequence 2.25 times the expected length of the tract.

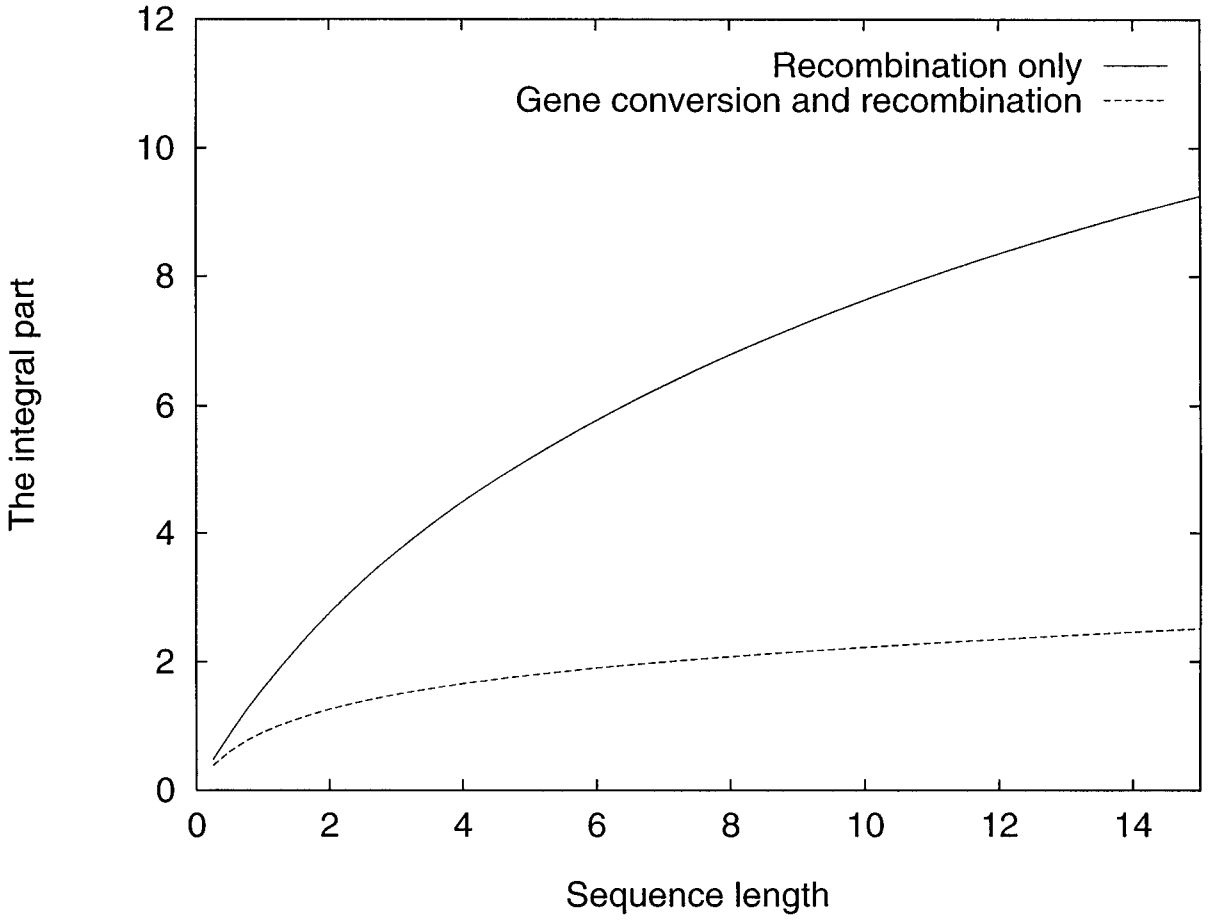

FIG. 5. The variance of the number of segregating sites, S_n. Shown is the integral part, $I(\lambda)$, of (28) as a function of sequence length λ with $n=2$ in a pure recombination model as well as in a model combining recombination and gene conversion. Sequence length is in expected number of recombinations per sequence per $4N$ generations. The variance depends on I through $\theta^2 I/2$, where θ is the rate of mutation in a sequence of length 1.

Drosophila melanogaster (see, e.g., Langley *et al.* (1999)) for an overview). In Fig. 4, the covariance between branch lengths, $B(\chi_1)$ and $B(\chi_2)$, in two positions, χ_1 and χ_2, is plotted as a function of the distance, λ, between the positions.

Hudson (1983) discusses the variance of the number of segregating sites, S_n, in a sample taken from the infinite-site coalescent model with recombination only. His result applies also to the infinite-site coalescent model with gene conversion and recombination,

$$\mathrm{Var}(S_n) = \lambda\theta \sum_{i=1}^{n-1} \frac{1}{i} + \frac{\theta^2}{2} \int_0^\lambda f_n(x)(\lambda - x)\, dx, \quad (28)$$

where $\theta = 4uNL$ is the mutation rate per sequence of length $\lambda = 1$ and $f_n(x)$ is the covariance between the total branch lengths of two trees x units away. The proof by Hudson (1983) carries over identically in this example. If $n=2, f_n(x)$ is given by (20) with r_λ replaced by ρ_x and for $n>2$ an approximation to $f_n(x)$ is given by (21) with r_λ replaced by ρ_x. In Fig. 5 the integral in (28) is plotted for $n=2$ both in a pure recombination model and in a combined model with the parameters obtained from the *Drosophila melanogaster* data.

DISCUSSION

We have developed and discussed a gene conversion model within the coalescent framework. General properties of the model have been worked out in terms of the distribution of the length of a gene conversion only.

Some implications of the model are proven, showing that results from the coalescent with recombination can be applied with minor modifications in several situations.

Further, it is shown that in *Drosophila melanogaster* the correlation between trees relating nearby positions is

significantly lower in a model that takes both gene conversion and recombination into account compared to a pure recombination model. This might account for the lack of associations between markers observed in *Drosophila melanogaster*, and it is in accordance with the prediction by Andolfatto and Nordborg (1997).

It is of interest to model gene conversion in a framework of various forms of selection. As discussed in Langley *et al.* (1999), both background selection and "hitchhiking" have been proposed as explanatory factors in *Drosophila* data. In the case of weak selection the gene conversion and recombination model developed here can easily be combined with the coalescent model with selection (Krone and Neuhauser, 1997).

The model can easily be extended to cover cases with variable populations sizes (see, e.g., Griffiths and Tavaré, 1994). Sample schemes and recursion formulas analogous to those found in Griffiths and Tavaré (1994) and Griffiths and Marjoram (1996) could be developed. Thus, a sample of observed sequences might be analyzed and inferences drawn as to G and parameters describing the distribution of the tract length.

ACKNOWLEDGMENTS

J. Hein is thanked for very valuable discussions and commenting on the manuscript. M. Nordborg is thanked for helpful comments and commenting on an early version of the paper and M. Schierup for a number of suggestions that improved the manuscript in various respects. The author was supported by Grant BBSRC 43/MMI09788 and by the Carlsberg Foundation, Denmark. Part of this work was carried out while the author visited the Isaac Newton Institute, University of Cambridge.

REFERENCES

Andolfatto, P., and Nordborg, M. 1997. The effect of gene conversion of intralocus associations, *Genetics* **148**, 1397–1399.

Begun, D. J., and Aquadro, C. F. 1995. Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*, *Mol. Biol. Evol.* **12**, 382–390.

Betrán, E., Rozas, J., Navarro, A., and Barbadillo, A. 1997. The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data, *Genetics* **146**, 89–99.

Griffiths, R. C. 1991. The two-locus ancestral graph, *in* "Selected Proceedings of the Symposium of Applied Probability, Sheffield 1989" (I. V. Basawa and R. L. Taylor, Eds.), IMS Lecture Notes—Monograph Series, Vol. 18, Inst. Math. Statist., Hayward, California.

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

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

Griffiths, R. C., and Tavaré, S. 1994. Sampling theory for neutral alleles in a varying environment, *Phil. Trans. R. Soc. London B* **344**, 403–410.

Hilliker, A. J., Harauz, G., Reaume, A. G., Gray, M., Clark, S. H., and Chovnick, A. 1994. Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*, *Genetics* **137**, 1019–1026.

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

Hudson, R. R. 1994. Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation, *J. Evol. Biol.* **7**, 535–548.

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

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

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

Krone, S. M., and Neuhauser, C. 1997. Ancestral processes with selection, *Theor. Popul. Biol.* **51**, 210–237.

Langley, C. H., Lazzaro, B. P., Phillips, W., Heikkinen, E., and Braverman, J. 1999. Linkage disequilibria and the site frequency spectra in the *su*(s) and *su*(w^a) regions of the *Drosophila melanogaster X* chromosome, submitted for publication, *available at* http://billy.ucdavis.edu/cemb.3.2 + tables.pdf.

Ohta, T. 1986. Actual number of alleles contained in a multigene family, *Genet. Res.* **48**, 119–123.

Stahl, F. W. 1994. The Holliday junction on its thirtieth anniversary, *Genetics* **138**, 241–246.

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

Wiuf, C., and Hein, J. 2000. The coalescent with gene conversion, in press.