The Coalescent With Gene Conversion

Carsten Wiuf* and Jotun Hein[†]

**Department of Statistics, University of Oxford, Oxford OX1 3TG, England and [†]Institute of Biological Sciences, University of Aarhus, 8000 Aarhus C, Denmark*

Manuscript received August 4, 1999
Accepted for publication January 18, 2000

ABSTRACT

In this article we develop a coalescent model with intralocus gene conversion. The distribution of the tract length is geometric in concordance with results published in the literature. We derive a simulation scheme and deduce a number of analytical results for this coalescent with gene conversion. We compare patterns of variability in samples simulated according to the coalescent with recombination with similar patterns simulated according to the coalescent with gene conversion alone. Further, an expression for the expected number of topology shifts in a sample of present-day sequences caused by gene conversion events is derived.

UNDERSTANDING patterns of variation in DNA sequences requires insight into at least three processes. The first is the genealogical process describing the ancestral history of a sample of alleles from a single locus. Kingman's coalescent process (Kingman 1982) is one such process. The second is the substitution process that describes how alleles mutate over time. For neutral loci the genealogical process and the substitution process can be separated such that the substitution process is superimposed on the genealogical process (Tavaré 1984), whereas for loci subject to selection the two processes cannot be separated. Finally, the last process describes the linkage relations between adjacent loci. It governs how and when adjacent loci are linked or split on different ancestral chromosomes and requires essentially an understanding of mechanisms that operate at the sequence level.

Central to many models describing the mechanisms at the sequence level is the Holliday junction (Holliday 1964). When the Holliday junction is resolved the result can be of two types: (1) a gene conversion with accompanying exchange of flanking regions (gene conversion with recombination); or (2) a gene conversion alone (Carpenter 1984; Stahl 1994). In this case a short tract of nucleotides is exchanged between two strands. We refer to the latter case as a gene conversion and the former as a recombination.

It is the aim of this article to develop a coalescent model with intralocus gene conversion alone and investigate the effects of gene conversion on various statistics of interest in the analysis of DNA sequences. We compare patterns of variability in samples simulated according to the coalescent with recombination with similar patterns simulated according to the coalescent with gene conversion alone. In Hilliker *et al.* (1994) the distribution of the gene conversion tract in *Drosophila melanogaster* is estimated. They find that a geometric distribution is a good approximation of the empirical distribution and accordingly the model is built on this observation.

In a subsequent article (Wiuf 2000) it is shown that to some extent the results given in this article apply generally without assuming any specific form of the distribution of the transferred chunk. The benefit of the geometric distribution is that more analytical results can be proven and that a specific choice of distribution (here, geometric) allows for investigation by simulation.

Two nucleotides separated by a large distance produce recombinants due to recombination events only. The (short) finite length of a gene conversion tract makes the contribution from gene conversion events insignificant. Thus, the patterns observed in a sample of sequences are the result of recombination events in the history of the sample and of the substitution process imposed. However, over small distances recombinants can be produced by recombination events as well as by gene conversion events. The observed patterns in samples of sequence data are thus the results of three different processes: the gene conversion process and the recombination process, and as well as the substitution process.

A recombination process was incorporated into Kingman's 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).

We expect that the patterns of intralocus variability in a model with gene conversion are different from

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

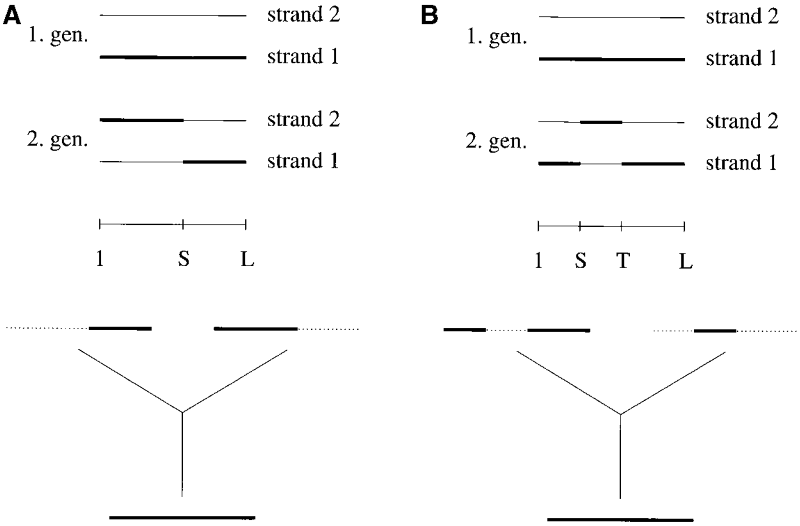

Figure 1.—(A) Recombination and (B) gene conversion. If the resolution of the Holliday junction results in an exchange of flanking regions we have (what here is called) a recombination event. In the top, two of the four strands involved in the Holliday junction are shown. In the bottom, time starts at present (the second generation) and goes backward. 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. If the Holliday junction is resolved without an exchange of flanking regions we have a gene conversion event (without recombination). In the top, two of the four strands involved in the Holliday junction are shown. In the bottom, time starts at present (the second generation) and goes backward. 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 between *S* and *T* shares another ancestor.

those in a model with recombinants produced by recombination (as defined above) only. The effect of recombination in the coalescent model is to break up the material ancestral to a sequence in two parts and distribute the parts on 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* (Figure 1). In contrast, gene conversion, as defined here, breaks the material ancestral to a sequence in two points, *S* and *T*, and distributes the material to the left of *S* and that to the right of *T* on one ancestor, and the part in between *S* and *T* is on another ancestor (Figure 1).

The effect of a recombination event can easily be obtained in a model of intralocus gene conversion. If one end point falls outside the observed sequence the effect will be similar to that of a recombination event, though the probabilities of the two events might be different from each other. These probabilities depend on the rates of gene conversion and of recombination within the observed sequence. Further, they depend on the number of nucleotides observed; the higher the number the less is the chance of a gene conversion with only one end point within the observed sequence.

Similarly, the effect of a gene conversion event can be obtained by two recombination events and one coalescent event (Figure 2). Again, the probabilities of obtaining the events might be very different. Especially, the latter series of events (given that the first recombination event occurs) will depend strongly on the current sample size; the higher the sample size, the lower the chance that the two recombined sequences will coalesce before coalescing with any other sequence in the sample.

Finally, the length distribution of the transferred chunk in the gene conversion events determines the distributions of the end points. Thus, the end points will only in rare cases be uniformly distributed along the observed sequence. Again, this is in striking discrepancy with the coalescent model with recombination where the breakpoint is uniformly distributed along the entire sequence (standard assumption; see, *e.g.*, Griffiths and Marjoram 1997).

Figure 3 shows an example of a genealogy of a sample of size 2 with intralocus gene conversion.

This article is organized into sections. The first two sections describe the coalescent model with gene conversion. In the third section a simulation scheme is developed similar to that of Griffiths and Marjoram (1997) to simulate from the coalescent with recombination; in the fourth section a number of analytical and

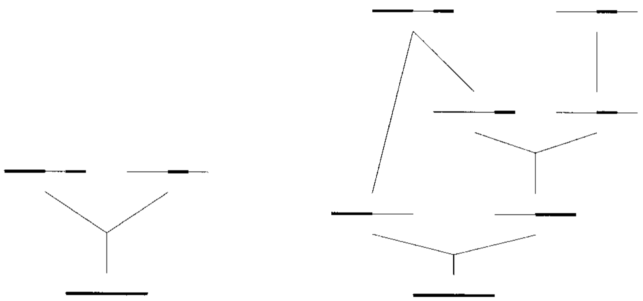

Figure 2.—Gene conversion *vs.* recombination. The topological effect of one gene conversion event (left) can be obtained in a model with recombination only by two recombination events accompanied by a coalescence event. Assuming the rate of gene conversion events is the same as the rate of recombination 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 it will result only in the desired distribution of ancestral material in one out of three possible mergings.

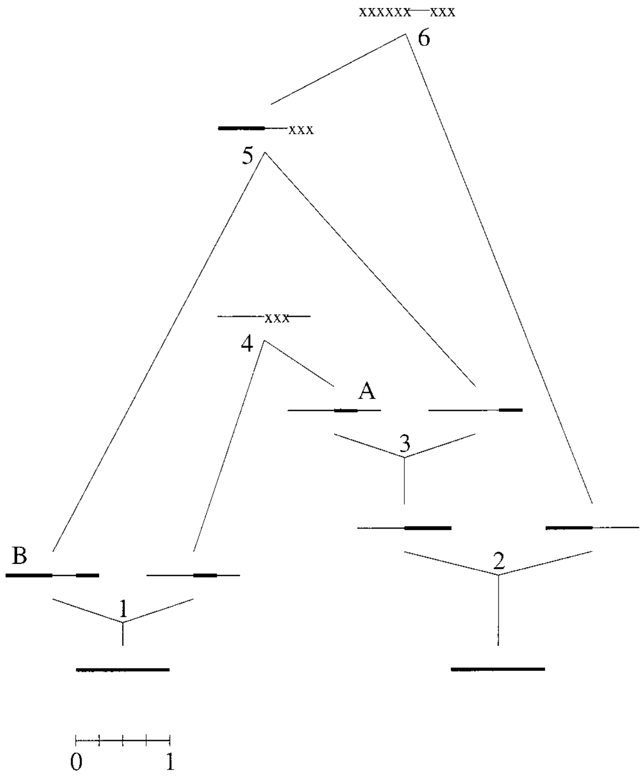

Figure 3.—Example of a sample history with intralocus gene conversion. We trace the history of a sample of size 2 back in time until all positions on the two sequences have found a most recent common ancestor (MRCA). We adopt the following graphical notation: thin lines represent material nonancestral to the present-day sample; thick lines, material ancestral to the sample; and crosses, material that has found a MRCA. The first three events, 1, 2, and 3, affecting the history of the sample are all gene conversion events spreading the material ancestral to the sample on five sequences. From the point of the topological structure of the tree, we cannot distinguish events 2 and 3 from recombination events. In the first event both end points of the transferred chunk fall within the observed sequence, whereas we cannot say whether the other end points in events 2 and 3 fall within the observed sequence length or fall outside. All that can be said is that the other end points do not fall within material ancestral to the present-day sample. The fourth event is a coalsecence event whereby positions (1/2, 3/4) find a MRCA. At event 5, positions (3/4, 1) find a MRCA, and at event 6, positions (0, 1/2) find a MRCA. Consider the ancestral sequence *A*. A gene conversion event with end points in the material not ancestral to the sample [that is outside (1/2, 3/4)] cannot be traced; it does not affect the history. Similarly, a gene conversion event in the ancestral sequence *B* with one or two end points in (1/2, 3/4) and none in ancestral material does not affect the history. Only the region spanned by ancestral material need be taken into account; *e.g.*, the span of ancestral material at sequence *B* is 1 − 0 = 1, and the span at sequence *A* is 0.75 − 0.5 = 0.25. Note that the region might include nonancestral material as in sequence *B*.

simulated results are collected. Results obtained in the coalescent with recombination apply on several occasions to the gene conversion model. Finally, the last section is a discussion of extensions to more realistic models of gene conversion and recombination. When proofs are omitted they can be derived from the general results in Wiuf (2000).

THE MODEL

Consider a diploid population model with effective constant (diploid) size $2N$. A new generation is obtained from the present generation by sampling $2N$ sequences with replacement, forming random pairs of sequences, and letting a short tract of nucleotides be transferred between the sequences. The mode of transfer is described below.

Sequences are of length $L + 1$ nucleotides so that there are L gaps between nucleotides. Consider one such sequence. Assume that in any generation the probability of a gene conversion initiating between any two positions in the sequence is g, independently of whether gene conversions initiate elsewhere along or outside the observed sequence. Where along the sequence a gene conversion initiates is uniformly distributed among all sites. If gL is small the chance of more than one gene conversion event in one generation is negligible.

The transferred chunk of nucleotides originates from a randomly chosen sequence in the population. In Wiuf (2000), the length, Z, of the converted chunk can have an arbitrary distribution. Here we consider the specific case where Z follows a geometric distribution with parameter q, *i.e.*,

$$P(Z|\text{conversion}) = q(1-q)^{Z-1}, \quad Z \geqslant 1. \tag{1}$$

That is, at least one nucleotide is transferred. It is assumed that the insertion happens to the right of the gap where the conversion initiates. This model is essentially equivalent to a model where the insertion happens to the right of the gap with probability p and to the left with probability $1 - p$ (see Wiuf 2000). The setting chosen is of mathematical convenience. If the conversion initiates at gap S, $S = 1, 2 \ldots, L$ (S for start), then the end point of the conversion is $T = S + Z$ (T for terminate). If $S + Z$ is $>L$, the conversion is partly outside the $L + 1$ observed nucleotides. The geometric distribution is in concordance with results in Hilliker *et al.* (1994) and emerges as a natural choice if the probability of adding an extra nucleotide to the transferred chunk does not depend on the current length. Later, we comment on other more realistic length distributions.

Let $Q = qL$. If q is small and L is large the distribution of $z = Z/L$ can be approximated with an exponential distribution with parameter Q, *i.e.*, $z \sim \text{Exp}(Q)$ (where $\sim$ means "is distributed like"). Further, let $G = 4gNL$ be the expected number of gene conversion events per sequence per $4N$ generations and assume g is small and N large. The definition of G is analogous to the parametrization adopted for the coalescence with recombination (Hudson 1983).

The genealogical process of a sample of n present-day sequences is studied: time starts at the present and increases, going backward in time. Under the above assumptions we find that the waiting time, W_C (in units of $2N$ generations), until a sequence has been created by a gene conversion event that initiates within the sequence is approximately exponentially distributed with parameter $G/2$, *i.e.*, $W_C \sim \text{Exp}(G/2)$, if N is large and gL is small. The rate of gene conversions initiating outside the sequence but ending within the observed sequence must also be taken into account. This rate depends on the distribution of Z and is discussed in the subsequent section. The rate of coalescence is $n(n - 1)/2$ if there are n sequences in a sample (Kingman 1982) assuming a constant population size.

We note that Q as well as G scale linearly in L, the observed number of gaps between nucleotides. That is, both Q and G are doubled if the observed number of gaps, L, is doubled. The expected length of the transferred chunk in a gene conversion is $1/q$ (measured in number of nucleotides). Therefore, the parameter Q is interpreted as the sequence length measured in units of expected length of the transferred chunk. Similarly, the parameter G is sequence length measured in expected number of gene conversion events per sequence per $4N$ generations. We note that $Q/G = q/(4gN)$ is independent of L.

EFFECTS OF GENE CONVERSION

In this section we discuss the effects of a single gene conversion event on a sequence and find the waiting time until the sequence has been created by a gene conversion event.

Denote by ζ, σ, and τ the variables Z/L, S/L, and T/L, respectively, where Z, S, and T are as defined in the previous section. The variables σ and τ take values in $\{1/L, 2/L, \ldots, 1 - 1/L\} \subseteq (0, 1)$ and as L becomes large the distributions of σ and τ converge to continuous distributions on $(0, 1)$. The representation of a sequence as the continuous interval $(0, 1)$ is commonly used (see, *e.g.*, Griffiths and Marjoram 1997).

Let C denote the event that a gene conversion happens in a given sequence 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, $S + Z > L$. Similarly, let C_2 denote that both end points are within the sequence length (lower index i indicates that i of the two end points of the converted chunk are within the L nucleotides). We find

$$P(C_2|C) = 1 - \frac{1}{Q}\{1 - \exp(-Q)\} \equiv K(Q) \qquad (2)$$

and

$$P(C_1|C) = 1 - P(C_2|C) = \frac{1}{Q}\{1 - \exp(-Q)\} = 1 - K(Q). \qquad (3)$$

Denote by W_{C_1} the time until an event of type C_1 and by W_{C_2} the time until an event of type C_2. Then we have $W_{C_1} \sim \text{Exp}(G(1 - K(Q))/2)$ and $W_{C_2} \sim \text{Exp}(GK(Q)/2)$, such that the waiting time, W_C, until a gene conversion event of either type is $W_C = \min(W_{C_1}, W_{C_2}) \sim \text{Exp}(G/2)$. This is similar to the recombination model; the waiting time to a recombination event within the observed sequence is $\text{Exp}(R/2)$, where $R = 4rNL$ is the rate of recombination and r is the probability of a recombination break between any two nucleotides in the sequence.

The end points of the tract are affected by the type, C_1 or C_2, of the event. The density, $f_{\sigma,\tau}(s, t|C_2)$, of σ and τ conditional on C_2 is given by

$$f_{\sigma,\tau}(s, t|C_2) = Q \exp(-Q(t - s))K^{-1}(Q), \quad 0 < s < t < 1, \qquad (4)$$

from which the marginal densities of σ and τ easily can be derived. The distribution of σ (and τ) is not uniform but skewed toward 0 (and 1). The density (4) is symmetric so that $(1 - \tau, 1 - \sigma)$ is distributed like (σ, τ).

The density of (σ, τ) conditional on C_2 relates to that of ζ conditional on C_2 because $\zeta = \tau - \sigma$. We find

$$f_\zeta(z|C_2) = Q(1 - z)\exp(-Qz)K^{-1}(Q), \quad 0 < z < 1, \qquad (5)$$

where $f_\zeta(z|C_2)$ denotes the density of ζ conditional on C_2.

Finally, the density of σ conditional on C_1 (that is, $\sigma + \zeta$ falls outside the observed nucleotides, $\sigma + \zeta > 1$) is

$$f_\sigma(s|C_1) = \exp(-Q(1 - s))(1 - K(Q))^{-1}, \quad 0 < s < 1. \qquad (6)$$

If the tract falls outside the observed sequence the chance that σ is close to 1 is higher than the chance that σ is close to 0. This is in concordance with (6); the density function is increasing in s.

In Figure 4 we illustrate the above different distributions with q obtained from *D. melanogaster* data.

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.

For most organisms, there is an upper bound to the length of a gene conversion tract. This means that only a minor (finite) extension of the observed sequence should be taken into consideration and not the entire chromosome. In the model discussed here, tract lengths can have an arbitrary size, but it can be shown that the probability of a tract initiating at least nL nucleotides away from the observed sequence and ending within the sequence, given a gene conversion event happens that ends in the observed sequence, is of order $\exp(-nQ)$.

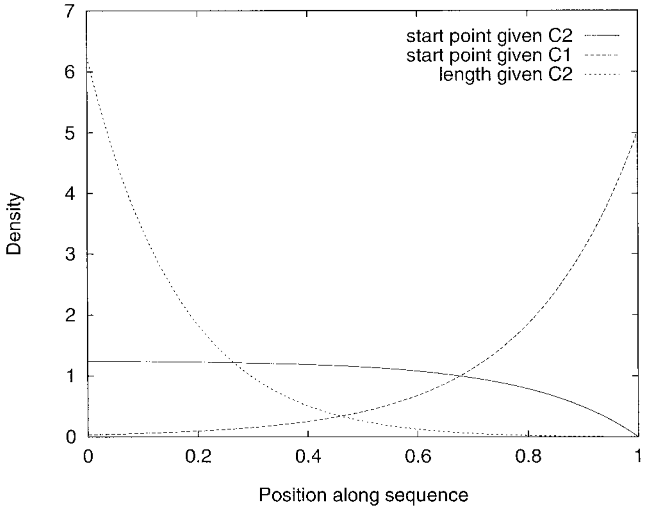

Figure 4.—Densities of the initiation point and the length of the gene conversion. Shown are the densities of σ conditional on C_1 and C_2, respectively, as well as the density of ζ conditional on C_2. The density of σ given C_2 is found from (4); $f_\sigma(s|C_2) = \{1 - \exp(-Q(1 - s))\}K^{-1}(Q)$, $0 < s < 1$. The parameter Q is in all graphs 5. Hilliker *et al.* (1994) estimates q in *D. melanogaster* to be ~1/350 and this corresponds in this case to an observed sequence length five times the expected length of a gene conversion tract or 1750 nucleotides, $L = Q/q$.

Basically only tracts initiating close to the observed sequence affect the sequence history in concordance with what is expected.

For each gene conversion tract initiating within the observed sequence and ending outside the observed sequence there is a similar tract initiating outside (to the left of the observed sequence), ending within the observed sequence. Because the chance of gene conversion events is the same along the entire chromosome, the two events have the same probability of happening.

Let C_o(o for outside) denote the event that a gene conversion initiates outside the observed sequence, and terminates within. The above informal argument shows that

$$P(C_o) = P(C_1|C) \tag{7}$$

and that the waiting time, W_{C_o}, until a gene conversion initiating outside the observed sequence ends within the observed sequence, is distributed like the waiting time, W_{C_1}, until an event of type C_1; that is, $W_{C_o} \sim W_{C_1}$. This is a general result that applies to the models described in Wiuf (2000).

Further, conditional on C_o, the termination point $t = T/L$ has density

$$\begin{aligned} f_\tau(t|C_o) &\approx \frac{Q\exp(-Qt)}{1 - \exp(-Q)} \\ &= \exp(-Qt)(1 - K(Q))^{-1}, \quad 0 < t < 1, \end{aligned} \tag{8}$$

because the length is exponentially distributed, and the chance that an exponential variable Exp(Q) is <1 is $1 - \exp(-Q)$.

The events and 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 approximately exponentially distributed with parameter $G(1 - K(Q))$, and where the initiation/termination point, δ, happens, is distributed with density

$$\begin{aligned} f_\sigma(s|C_o \cup C_1) = \frac{1}{2}\{\exp(-Q(1 - s)) + \exp(-Qs)\} \\ \cdot (1 - K(Q))^{-1}, \quad 0 < s < 1, \end{aligned} \tag{9}$$

according to (6) and (8).

Summing up, we find that the waiting time, W, until a sequence is created by a gene conversion event is exponentially distributed

$$\begin{aligned} W &\sim \operatorname{Exp}\left(\frac{G}{2}\left\{1 + \frac{1}{Q}(1 - \exp(-Q))\right\}\right) \\ &\sim \operatorname{Exp}\left(\frac{G}{2}\{2 - K(Q)\}\right), \end{aligned} \tag{10}$$

because $P(C_o) + P(C_1|C) + P(C_2|C) = 2 - K(Q)$.

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

$$\begin{aligned} p_2 &= \frac{1 - (1/Q)(1 - \exp(-Q))}{1 + (1/Q)(1 - \exp(-Q))} \\ &= \frac{K(Q)}{2 - K(Q)}, \end{aligned} \tag{11}$$

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

$$\begin{aligned} p_1 &= \frac{(2/Q)(1 - \exp(-Q))}{1 + (1/Q)(1 - \exp(-Q))} \\ &= \frac{2(1 - K(Q))}{2 - K(Q)}. \end{aligned} \tag{12}$$

Whether it is a type C_o or type C_1 event happens with probability 1/2. The density of (σ, τ) is given by (4) (if there are two breakpoints) and the density of the single breakpoint σ is in the last case given by (9). In the example in Figure 4, $Q = 5$, or five times the expected tract length, and we find $p_2 = 0.67$ and $p_1 = 0.33$.

The distribution in (10) is interesting for several reasons. First, the distribution of the tract length, ζ, depending on Q, affects the number of events and the times between events. The parameter $G(2 - K(Q))/2$ varies from $G/2$ when $Q = \infty$ to G when $Q = 0$. Second, for all $Q > 0$ the number of events is higher than the number of events in a recombination model with a similar parameter (*i.e.*, $r = g$).

The two extreme values, $Q = 0$ and $Q = \infty$, are of

```
Put $k = n$, and $i = j = 1$,
While $k > 1$,
    Choose $W_i$ exponential with parameter $k(k-1)/2 + k\tilde{G}/2$,
    If $U_i < (k-1)/(k-1+\tilde{G})$,
        Choose two sequences at random and let them coalesce,
        Put $k = k - 1$,
    Else
        Choose one sequence at random,
        If $V_j < p_1$,
            Choose one point on the chosen sequence according
            to (4), and form two sequences,
        Else
            Choose two points on the sequence according to (9),
            and form two sequences,
        Put $k = k + 1$, and $j = j + 1$,
    Put $i = i + 1$,
End.
```

Figure 5.—The simulation scheme. Define $\tilde{G} = G(1 + (1 - \exp(-Q))/Q) = G(2 - K(Q))$, and let k denote the number of sequences ancestral to the sample of size n at a given time in the past. Further, let U_i, $i \geq 1$, and V_j, $j \geq 1$, be series of uniform variables on (0, 1). The exponential variable W_i is the time between event $i - 1$ and i in the samples history. Events are either coalescence or gene conversions, and if a gene conversion occurs we distinguish between two types according to the number of end points within the observed sequence. With probability p_1 there is one end point only, otherwise two. The index i counts the number of events and j the number of gene conversions. That is, $i - j$ is the number of coalescence events. The algorithm stops the first time there is one ancestor to the sample.

interest. If $Q = 0$ we find that $W \sim \text{Exp}(G)$, and a gene conversion will leave just one breakpoint within the sequence. Where the break occurs will be uniformly distributed along the sequence. Thus, the coalescent with gene conversion resembles the coalescent with recombination, but the rate of gene conversions is twice the rate of recombinations (assuming the probability of a recombination between any two nucleotides is g).

If $Q = \infty$ we find that $W \sim \text{Exp}(G/2)$ and that both end points of a gene conversion will be within the sequence. As Q goes toward infinity, σ will be uniformly distributed along the sequence, and $\tau \approx \sigma$, that is, $\zeta \approx 0$. In the limit $Q = \infty$, a gene conversion will just leave a spot on the sequence, leaving the sequence the way it is. Thus, the coalescent with gene conversion will be indistinguishable from the pure coalescent process. Each time a conversion happens, it cannot be seen, and only coalescent events affect the history of a sample.

MODEL SIMULATIONS

Assume n sequences are sampled from a present-day population. The genealogy of the sample subject to gene conversion as described in the previous section can be simulated according to the scheme in Figure 5.

The algorithm is formulated as a birth and death process with constant death rate, $G(2 - K((Q))/2$, and terminates when there is only one ancestral sequence to the sample. Griffiths and Marjoram (1997) developed a similar birth and death process to handle to coalescent with recombination. Both algorithms return a genealogy in which the "true" genealogy of the sample is embedded; we keep track only of the number of sequences and the end points of gene conversions, not which parts of the sequences that are ancestral to the sample. Gene conversions can happen outside the ancestral material to a sequence, thereby creating an "empty" sequence, *i.e.*, a sequence with no material ancestral to the sample (*cf.* Figure 2).

Consider a fixed point, χ, along the sequences. The tree, $T(\chi)$, that describes the sequences at χ can be found by starting at the present sequences and going back in time. When a gene conversion node is encountered, the branch that describes the segment containing χ is followed. Assume that the gene conversion leaves only one end point, s, in the sequence. If $s < \chi$, then the branch describing the fate of the left part of the sequences is to be followed and vice versa and similarly, if the gene conversion leaves two end points. The tree $T(\chi)$ is distributed like the coalescent process since one point cannot be subject to gene conversion.

Mutations can be superimposed on the genealogy if all alleles are selectively neutral. Assuming mutations arrive according to a Poisson process, the number of mutations on the total genealogy is Poisson with parameter $\theta B/2$, where B is the total branch length of the entire genealogy, $\theta = 4Nu$ is the scaled mutation rate, and u is the probability of a mutation in a sequence per generation.

The coalescent with gene conversion can be combined with the coalescent with recombination. The genealogy of a sample of sequences is constructed similarly to the scheme above, waiting for three different events to occur: coalescence, recombinations, and gene conversions.

RESULTS

As noted previously, $Q_0 = Q/G = q/(4gN)$ is independent of L. The parameters (Q_0, G) are more natural parameters than (Q, G); G represents the length of the sequence in units of gene conversion events per sequence per $4N$ generations and Q_0 is a parameter dependent on the effective size of the population and parameters intrinsic to the biological system. In contrast (Q, G) are both parameters representing sequence length, but in different units.

In the following, results are given in terms of (Q_0, G) to facilitate comparisons between samples of different lengths but with the same Q_0. If we consider Q_0 fixed and G a variable, results (and plots) can be converted between models with different g's but equal Q_0 through a linear scaling of sequence length.

If sequence length is measured in units of expected

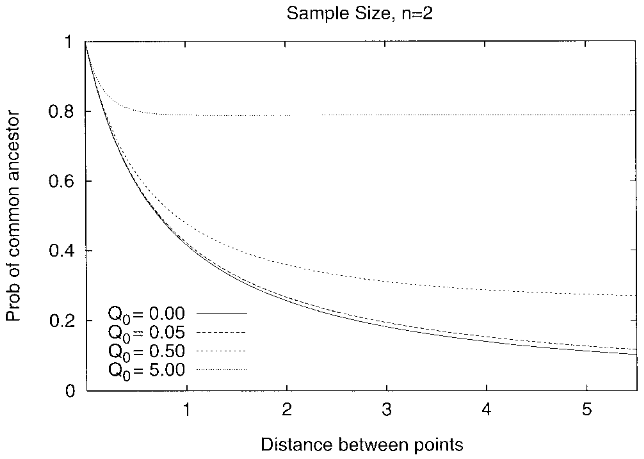

Figure 6.—Probability that two positions share an ancestor, $n = 2$. Shown is equation 15 for different values of Q_0. Sequence length is in expected number of gene conversions per sequence per $4N$ generations. Using estimates in Hilliker *et al.* (1994), we find $Q_0 = 0.05$ in *D. melanogaster* ($g = 3 \times 10^{-8}$, $q = 3 \times 10^{-3}$, and $2N = 10^6$). The curve for $Q_0 = 0.05$ flattens out approximately at 0.027 for $G > 40$ in which case $Q > 2$, or twice the expected length of a gene conversion tract. The case $Q_0 = 0$ corresponds to a model with recombination only at rate $2G$.

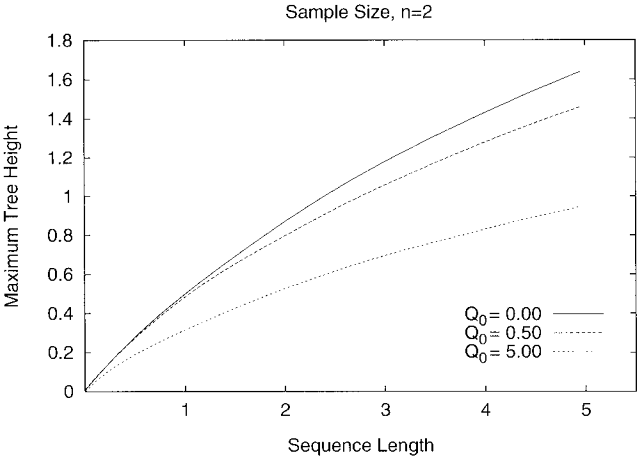

Figure 7.—Simulated values of the time until all positions have found a MRCA in a sample of size $n = 2$. The expected value $2(1 - 1/n)$ of the time to a MRCA in a single position is subtracted. The curves increase with increasing values of Q_0. This is in agreement with the fact that the expected number of gene conversions also increases with increasing Q_0 (see Figure 9). The *D. melanogaster* curve, $Q_0 = 0.05$, follows $Q_0 = 0$ very closely (not shown). A total of 10,000 simulations were performed for each value of Q_0.

number of events per sequence per $4N$ generations, we find that the scaled length $\zeta' = \zeta G$ of a tract is

$$f_{\zeta'}(z) = Q_0 \exp(-zQ_0), \quad z > 0 \tag{13}$$

[according to (1)]; that is, exponential with parameter $Q_0 = q/(4Ng)$. The waiting time, W, until a sequence is created by a gene conversion event is in this formulation given by

$$W \sim \mathrm{Exp}\left(\frac{G}{2} + \frac{1}{2Q_0}(1 - \exp(-Q_0 G))\right) \tag{14}$$

[from (10)]. The probabilities of the different types of gene conversion events are given by (11) and (12) with Q replaced by Q_0G. A similar remark applies to the densities of the end points σ and τ.

In the following, we consider a sample of size n taken from the population at the present time. First, we focus on correlations between trees. Consider two positions, χ_1 and χ_2, in a sample of n sequences at distance G. 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 type 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, $r_G/2$, by which events of type C_1 and C_o happen is, per sequence,

$$r_G = 2\,\frac{1}{Q_0}(1 - \exp(-Q_0 G));$$

cf. (10) and (12).

For small values of G we find $r_G \approx 2G$. This is of particular interest because the rate at which recombinants are produced by recombination is only G (assuming $r = g$), and this is half the rate at which recombinants are produced by gene conversion events. Thus, gene conversion events might contribute significantly to linkage disequilibrium over small distances (see also Wiuf 1999). For example, according to Andolfatto and Nordborg (1997), $g = 3r$ in *D. melanogaster* and the rate of gene conversions is thus sixfold that of recombination over small distances.

Applying results in Griffiths (1991), we find for $n = 2$,

$$\begin{aligned} P(T_2(\chi_1) = T_2(\chi_2)) &= \mathrm{Cov}(T_2(\chi_1), T_2(\chi_2)) \\ &= \frac{18 + r_G}{18 + 13r_G + r_G^2}, \end{aligned} \tag{15}$$

where Cov denotes the covariance between variables. We note that the expression in (15) obtains its minimum when $G = \infty$, in which case $r_G = 2/Q_0$. Thus the covariance is strictly positive for all values of G and converges toward a nonzero constant as $G \rightarrow \infty$. We have

$$P(T_2(\chi_1) = T_2(\chi_2)) \geqslant \frac{(9Q_0 + 1)Q_0}{9Q_0^2 + 13Q_0 + 2}. \tag{16}$$

The probability $P(T_2(\chi_1) = T_2(\chi_2))$ for $n = 2$ is plotted in Figure 6. The tree height, $T_n = \max\{T_n(\chi)|\chi\}$, until all positions in the sample have found a most recent common ancestor is plotted in Figure 7 for $n = 2$. For $Q_0 = 0.05$ and $G = 5$ we find $Q = 0.25$, which corresponds to ~100 nucleotides in *D. melanogaster* (see Figures 4 and 6). In this case the chance that two positions separated at distance 5 share a most recent common ancestor (MRCA) is ~0.10 if sample size is 2.

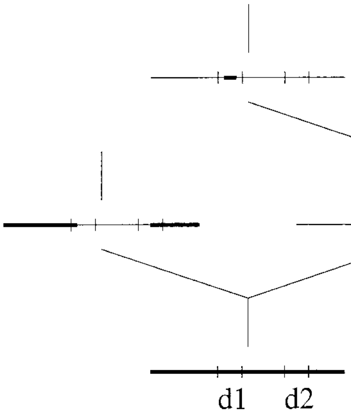

Figure 8.—Number of gene conversion events. The number of gene conversion events affecting the history cannot be counted by chopping up the sequence into small sections, d_i, $i = 1, 2, \ldots$, counting the number, c_i, in each section and then summing over i, $\Sigma_i c_i$. Knowing that one end point of a gene conversion is in section d_i does not provide information whether or not the other end point is within ancestral material. In the example, $c_1 = 2$, $c_2 = 1$, and $c_1 + c_2 = 3$, but only two events affect the history. Each breakpoint occurs twice in the drawing.

Also of interest are the numbers of different types of gene conversion events. We discuss two such numbers. The first is the number of gene conversion events affecting the history of the sample. Denote this number by G_n. The second is the number of gene conversion end points that make the topology change. Call this number S_n. This number relates to the possibilities of detecting gene conversion events in the sample history. Under an infinite-sites mutation model (Watterson 1975) with a mutation rate very high compared to that of gene conversions, G, all shifts in topologies can be detected from sequence polymorphisms (Hudson and Kaplan 1985). Also, these are the only events that can be detected (under the infinite-sites model).

It is not possible to find the expectation of G_n analytically. The reason for this is the following. The expectation of G_n is not linear in G because a gene conversion with both end points in ancestral material is only counted once. Knowing that a gene conversion has one end point within ancestral material in a small sequence interval does not reveal if the other end point also is within ancestral material. Thus, information about gene conversion end points in a small sequence interval does not provide information whether the points count in G_n (Figure 8). In the coalescent with recombination, the expected number of recombination events can easily be found because the number is linear in R (Hudson and Kaplan 1985). The expected value of G_n is simulated for different values of G and Q_0 and plotted in Figure 9. Asymptotically, the expectation of G_n is linear.

The expected number of events causing a topology shift is given by

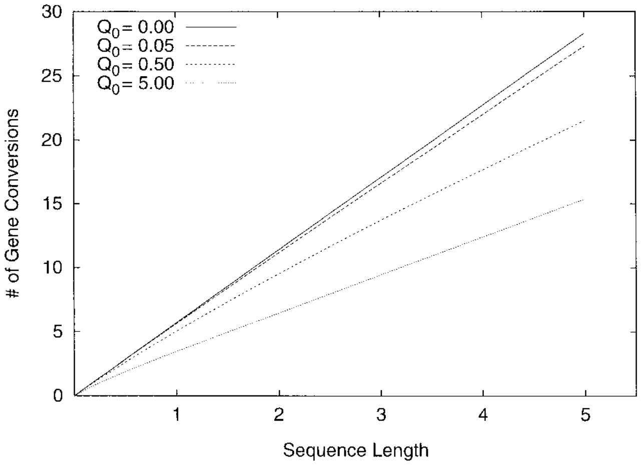

Figure 9.—Expected value of G_n. Sequence length is in expected number of gene conversions per sequence per $4N$ generations. As sequence length increases the expected number of gene conversions becomes effectively linear. The sample size is 10 and 10,000 simulations were performed for each value of Q_0.

$$E[S_n] \approx 2\left\{\sum_{i=1}^{n-1}\frac{1}{i} - 2.14\right\}G. \tag{17}$$

The proof is similar to that of Hudson and Kaplan (1985; see Wiuf *et al.* 1999 for the approximation). The difference is the factor 2 in front, which comes from the fact that $r_G \approx 2G$ for small G and not $r_G = G$ as with recombination. Again, this can result in a considerable difference in the number of topology shifts seen in models with recombination *vs.* gene conversion.

Consider position 0 in the sequences and let B_n be the total length of the coalescent tree in position 0. Wiuf and Hein (1999) consider the coalescent with recombination and discuss the sequence length, Λ_n, conditional on B_n, from position 0 until the first recombination point affecting the history of the sample. They find that

$$P(\Lambda_n > x|B_n = b) = \exp(-bx/2), \quad x > 0, \tag{18}$$

where sequence length is measured in units of expected number of recombination events per sequence per $4N$ generations. Sequences are here potentially infinitely long.

Here, we give the analogous result for the coalescent model with gene conversion. The situation differs in the sense that we wait for either a gene conversion initiating to the right of position 0 [analogous to (18)] or a gene conversion initiating to the left of position 0 but ending to the right of 0. The proof is given in the appendix. The distribution of the length, Λ_n, until the first point affecting the history of the sample, conditional on B_n, is

$$P(\Lambda_n > x|B_n = b) = \exp\{(e^{-Q_0x} - 1)b/(2Q_0) - bx/2\}, \quad x > 0. \tag{19}$$

Similarly, the distribution of Λ_n can be found (see appendix) and is given by

$$P(\Lambda_n > x) = \prod_{i=1}^{n-1} \frac{i}{i + x + (1 - \exp(-xQ_0))/Q_0}. \qquad (20)$$

If $n = 2$ we find that the expectation of Λ_n is infinite; that is, in general the sequence length until the first gene conversion event is large. For $n > 2$ the expectation is finite.

DISCUSSION

We developed a coalescent model with gene conversion assuming a diploid population of constant size, N. It takes two parameters as input (along with the sample size). The first is the product $G = 4NLg$, where g is the probability that a gene conversion tract initiates in a fixed position, and the second parameter is $Q = qL$, where q is the probability that the next nucleotide is within the tract given that the former is within the tract. An easy simulation scheme was developed. This scheme could be modified, in accordance with Griffiths and Tavaré (1994), to allow for fluctuation in the population size with time.

We derived a number of results related to the correlation of trees in different positions and to the probability that two given positions share a common ancestor. These tend to be highly different from similar quantities obtained in the coalescent with recombination. The covariance between trees relating two distinct positions does not tend to zero with increasing distance between the positions, but is bounded by a positive constant. Thus, it is expected that the trees in the two distinct positions are more likely to share the same topology only in a model of gene conversion than in a model of recombination with similar rate. Figure 10 confirms this. As $Q_0 = Q/G$ increases, the probability, π, of common topology in two distinct positions decreases and for fixed value of Q_0, π converges (as function of distance) to a level distinct from 0 and 1. Low values of Q_0 have a similar effect to that of recombination.

We should note in passing that given a gene conversion end point is in position χ, the probability, p, of detecting that the gene conversion has occurred from incompatibilities in the sequence alignment is very low and slowly increases with sample size. Hudson and Kaplan (1985) and C. Wiuf, T. Christensen and J. Hein (unpublished results) found

$$p \le 1 - \frac{2.14}{\log(n)}$$

for large sample sizes, n; *e.g.*, if $n = 1000$, $p \le 0.69$. The proof is similar to that of Hudson and Kaplan (1985) and relates to the expected number of topology shifts [see Hudson and Kaplan 1985 and formula (17)].

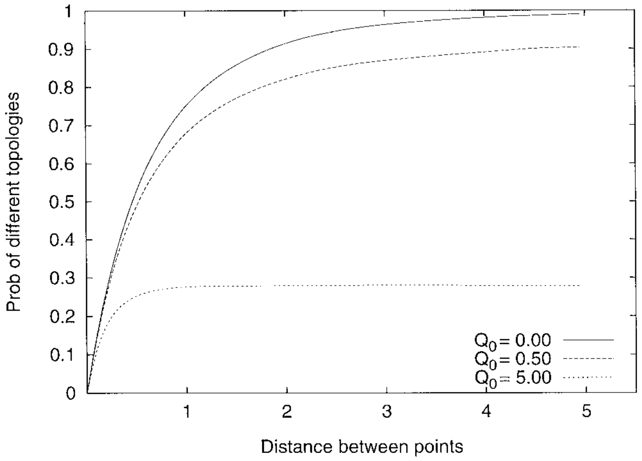

Figure 10.—Simulated probabilities that two positions do not share the same topology for various values of Q_0. Sequence length is in expected number of gene conversions per sequence per $4N$ generations. Consider again *D. melanogaster*. The curve $Q_0 = 0.95$ follows $Q_0 = 0$ very closely and the chance that two points at distance 1 have different topologies is >70%. This corresponds to a distance of 5% the length of the expected tract length or $0.05 \times 350 \approx 20$ nucleotides. Sample size is 10 and 20,000 simulations were performed for each value of Q_0.

It is of high interest to be able to distinguish mechanisms operating on the molecular level in patterns of variability obtained in a sample of sequences. An indicator of gene conversion could be the following. Define an informative site to be a site where at least two different nucleotides are present in at least two sequences each. Further, say a pair of sites is compatible if there exists a topology explaining the sites with the minimum possible number of substitution events. Now, assume three informative sites are given. If the first and last sites are compatible, and the two remaining pairs involving the middle site are both incompatible, we would take this as evidence of a gene conversion event. In a pure recombination model, given the middle site is incompatible with both other sites, we would expect that the first and the last sites are also incompatible. We have simulated how often under ideal circumstances this is likely to happen. It is assumed that the topology can be accurately constructed for each column in the sequence alignment. This is, for example, the case in an infinite-sites model with very high mutation rate. Table 1 shows simulation results for various values of Q_0 and G.

Denote the positions 1, 2, and 3. Positions 1 and 3 are fixed whereas 3 is uniformly distributed between 1 and 3. The findings in the table can be explained as follows. Assume $Q_0 = 0$. If 1 and 2 do not share topology and similarly for 2 and 3, this requires at least two recombination events. As the distance between 1 and 3 increases, the chance of more recombination events before the MRCAs to positions 1 and 3 increases; thus the chance that 1 and 3 share the same topology decreases.

If $Q_0 > 0$ the pattern changes. When the distance is

TABLE 1

Distance between 1 and 3

	Distance		
Q_0	1	5	10
0	3.4	0.35	0.037
0.05	3.8	0.74	0.12
0.5	10.5	6.21	6.31
5	60.8	63.2	—

Shown are simulated values of the probability in percentage that two positions, 1 and 3, share the same topology given positions 1 and 2, and 2 and 3 do not share topology. Position 2 is uniformly distributed between 1 and 3. A total of 20,000 simulations were performed for each entry. The value for $Q_0 =$ 5 and distance 10 not found due to computational difficulties.

small compared to $1/Q_0$ (expected tract length), most events will be of type C_1 and C_o. A similar pattern to that for $Q_0 = 0$ is thus expected. As the distance increases, the rate by which 1 and 3 are separated becomes almost constant ($r_G = 2\{1 - \exp(-Q_0G)\}/Q_0$) and if (1, 2) and (2, 3) do not share topologies it is most likely caused by a gene conversion event with both end points within 1 and 3. Thus we will accept more cases where 1 and 3 share topology.

Now fix the distance between the positions. As Q_0 increases, the number of C_2 events increases relative to C_1 and C_o events (Equations 11 and 12) and again if (1, 2) and (2, 3) do not share topologies it is most likely caused by a gene conversion event that removes a region around position 2. Again, we expect more cases where 1 and 3 share topology.

The pattern observed in Table 1 is obtained under ideal circumstances. In practice, the topology relating a set of sequences in a given position can be reconstructed from the nucleotide pattern in that position in rare cases only. Under an infinite-sites model with low mutation rate, the chance that 1 and 3 can be explained by the same topology is higher than the value given in Table 1. Similarly, recurrent substitutions have the same effect, but in both cases the overall pattern in Table 1 is retained. These results suggest that in a pure recombination model we will see many more pairs of incompatible sites than in a model with pure gene conversion (assuming the rate of recombination is similar to the rate of gene conversion). On the other hand we expect many more topology shifts under a gene conversion model, so that we must return, moving along the sequences, to previously visited topologies more often than under a recombination model. Also, the results in Table 1 imply that recombination can be distinguished from gene conversion by the spatial arrangement of incompatible pairs, simply by estimating the probability in Table 1 from data. If recombination is not present we will expect this probability to be significantly different from 0.

It would be advantageous to build a model of gene conversions and recombination more directly on a molecular model of these phenomena. This will be pursued in greater depth elsewhere, but the basic issues involved are shortly sketched here.

Central to most models of gene conversion / recombination is the Holliday junction (Holliday 1964). Although the original Holliday model has since been superceded by more complicated models, such as the single strand invasion model (Meselson and Radding 1975) and the double-chain-break repair model (Resnick 1976), the Holliday model would be still be the natural starting point for modeling gene conversion from a population genetical viewpoint. Issues in the Holliday model would also be central in the more complicated models. In the Holliday model the following three problems would be of central importance.

1. The movement of the Holliday junction and its consequences. If the movement is controlled by a random walk that is stopped at a fixed time, the length of the gene conversion would be close to a symmetric binominal distribution moved so its mean is zero. Since the movement of the Holliday junction along nonidentical sequences will create heteroduplexes that are energetically unfavorable to homoduplexes, it could be expected that there would be a sequence-dependent centralizing drift in this random walk. Since the energetic costs of mismatches are approximately known, this distribution could be calculated. The length of gene conversions would then be sequence dependent. This would be a complication when incorporating it into the coalescent that takes a view from the present toward the past.
2. How is the heteroduplex corrected? Two extremes can be imagined. The heteroduplex is resolved when the DNA is duplicated or one single strand is made the master copy and the other single strand is corrected so it matches to this master strand. From the point of view of redistributing ancestral material, this is simple as the region within the gene conversion will choose one ancestor. The other possibility, that mismatches are corrected independently one by one, has more drastic consequences for the distribution of ancestral material as each correction will be independent choices of parents. Within the region of the gene conversion, different positions can have different ancestors. This would also have drastic consequences for the phylogenetic relationships among the sequences as one moves along the sequences, as even arbitrarily close positions could be related by different phylogenies.
3. The Holliday structure can be resolved in two ways, one leading to a gene conversion, but no recombination, and the other leading to a gene conversion and a recombination. What are the probabilities of gene conversion and recombination? Few, if any, biologi-

> cal systems have gene conversion but no recombination. In modeling the consequences, it would be natural to make this probability a single parameter in the model.

Later models, especially Meselson and Radding (1975) and the double exchange model (Resnick 1976), are more complicated and will slightly change the picture. From a population genetical viewpoint, all that matters in these different models is how the ancestral material is redistributed when viewed from the present toward the past. The Meselson-Radding model would have a very short region with a one-directional gene conversion added relative to the pure Holliday model, due to the strand invasion. Since this region is expected to be short relative to the length of the migration of the Holliday structure, the Meselson-Radding model is expected to be very similar to the pure Holliday model. The double-chain-break repair model would have two strand invasions and also two Holliday junctions that could undergo a random walk relative to each other. The double-chain-break repair model probably has greater consequences for the distribution of ancestral material than the Meselson-Radding model, but that remains to be explored.

Since it is not well known how different recombination mechanisms are phylogenetically distributed, it would be interesting to know if it was ever possible to distinguish the underlying mechanism from pure population data. It would here be natural to formulate what was expected in terms of different sets of incompatibilities. For instance, if three informative sites were given, and the first and last sites were compatible, but the two remaining pairs involving the middle site were incompatible, this would be taken as an indicator of a gene conversion. If more informative sites were found close to the middle position, these would be expected to have different kinds of compatibilities dependent on which kind of repair mechanism was operating on the heteroduplex.

Whether or not different recombination/gene conversion models are distinguishable by population sequence data depends on the frequency of configurations of segregating nucleotides in the population that would result in different products when recombined under different models. A functional geneticist would design the necessary configurations of nucleotides and set up the adequate crosses. In population genetics this would have to occur by chance. It is unlikely that population sequence data can compete with designed experiments in this respect, but this does not imply that different mechanisms of recombination will not have consequences for the expected patterns of variation in population sequence data; such data are becoming available from many organisms, where molecular genetical experiments have not yet been performed.

J. P. Hjorth is thanked for helpful discussions on what gene conversion is, M. Nordborg for commenting on an early version of the manuscript, and M. Schierup for various suggestions that improved the manuscript considerably. We thank A. Mikkelsen for implementing the model and performing the simulations used in the article. C.W. was supported by grant Biotechnology and Biological Sciences Research Council 43/MMI09788 and by the Carlsberg Foundation, Denmark. Part of this work was carried out while the authors visited the Isaac Newton Institute, University of Cambridge.

LITERATURE CITED

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

Carpenter, A. T. C., 1984 Meiotic roles of crossing-over and of gene conversion. Cold Spring Harbor Symp. Quant. Biol. **49:** 23–29.

Griffiths, R. C., 1991 The two-locus ancestral graph, pp. 100–117 in *Selected Proceedings of the Symposium of Applied Probability, Sheffield 1989, IMS Lecture Notes-Monograph Series, 18*, edited by I. V. Basawa and R. L. Taylor. Hayward, CA.

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

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

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

Holliday, R., 1964 A mechanism for gene conversion in fungi. Genet. Res. **5:** 282–287.

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

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

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

Meselson, M. S., and C. M. Radding, 1975 A general model for genetic recombination. Proc. Natl. Acad. Sci. USA **41:** 215–220.

Resnick, M. A., 1976 The repair of double-strand breaks in DNAs: a model involving recombination. J. Theor. Biol. **59:** 97–106.

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

Tavaré, S., 1984 Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

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

Wiuf, C., 2000 A coalescent approach to gene conversion. Theor. Popul. Biol. (in press).

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

Wiuf, C., and J. Hein, 1999 Recombination as a point process along sequences. Theor. Popul. Biol. **55:** 248–259.

Communicating editor: A. G. Clark

APPENDIX: PROOF OF FORMULAS (19) AND (20)

Wiuf and Hein (1999) consider the coalescent with recombination and discuss the sequence length, Λ_n, conditional on the total branch length, B_n, from position 0 until the first recombination point affecting the history of the sample. They find that

$$P(\Lambda_n > x | B_n = b) = \exp(-bx), \quad x > 0, \qquad \text{(A1)}$$

where the sequence length is measured in units of expected number of recombinations per sequence per $4N$ generations. Sequences are here potentially infinitely long.

Remember [see (10), (11), and (12)]

$$W_{C_o} \sim \mathrm{Exp}(G(1 - K(Q_0G))/2)$$

and

$$W_{C_1 \cup C_2} \sim \mathrm{Exp}(Q_0G/2).$$

Whenever an event of type $C = C_1 \cup C_2$ happens the initiation point is uniform along the sequence (per definition). Therefore, the sequence length, $\Lambda_{1,2}$ from position 0 until the first breakpoint affecting the history of the sample is distributed like (A1)

$$P(\Lambda_{1,2} > x|B_n = b) = \exp(-bx/2), \quad x > 0, \tag{A2}$$

for sequences potentially infinitely long.

The rate, $G(1 - K(Q_0G))/2$, of type C_o events converges to $1/(2Q_0)$ for $G \to \infty$. Conditional on $B_n = b$, the number, G_o, of C_o events in the sequences ancestral to position 0 is thus Poisson distributed with parameter $b/(2Q_0)$,

$$P(G_o = k|B_n = b) = \frac{1}{k!}\left(\frac{b}{2Q_0}\right)^k \exp(-b/2Q_0)), \quad k = 0, 1, \ldots. \tag{A3}$$

The density, $f_\tau(x|C_o, G)$, of the termination point $\tau_G = G\tau$ (in units of sequence length) converges to an exponential distribution for $G \to \infty$

$$f_\tau(x|C_o, G) = Q_0 \exp(-Q_0x)\ (1 - \exp(-Q_0G))^{-1} \to Q_0 \exp(-Q_0x), \quad x > 0 \tag{A4}$$

[from (8)]. Combining (A3) and (A4) we find the distribution of the length, Λ_o, until the first point of type C_o affecting the history of the sample,

$$\begin{aligned} P\ (\Lambda_o > x|G_o > 0, B_n = b) &= \sum_{k=1}^{\infty} \exp(-kQ_0x)\frac{1}{k!}\left(\frac{b}{2Q_0}\right)^k \cdot \frac{\exp(-b/(2Q_0))}{1 - \exp(-b/(2Q_0))} \\ &= \frac{\exp(-b/(2Q_0))}{1 - \exp(-b/(2Q_0))} \cdot \{\exp(e^{-Q_0x}b/(2Q_0)) - 1\}, \quad x > 0. \end{aligned} \tag{A5}$$

Outside $\{G_0 > 0\}$, $\Lambda_0 = \infty$.

Finally, combining (A2), (A3), and (A5) we deduce that $\Lambda_n = \min\ (\Lambda_o, \Lambda_{1,2})$ has distribution

$$P(\Lambda_n > x|B = b_n) = \exp\{(e^{-Q_0x} - 1)b/(2Q_0) - bx/2\}, \quad x > 0, \tag{A6}$$

as required. The distribution of Λ_n can be found in the following way. The event $\{\Lambda_n > x\}$ is equivalent to the event that no gene conversion events occur in a sample of sequences of length x before a MRCA is found. This has probability

$$P(\Lambda_n > x) = \prod_{i=1}^{n-1} \frac{i}{i + x + (1 - \exp(-xQ_0))/Q_0}. \tag{A7}$$