Copyright © 2003 by the Genetics Society of America

A Coalescent Model of Recombination Hotspots

Carsten Wiuf[1] **and David Posada**

Variagenics, Cambridge, Massachusetts 02139

Manuscript received August 2, 2002
Accepted for publication February 3, 2003

ABSTRACT

Recent experimental findings suggest that the assumption of a homogeneous recombination rate along the human genome is too naive. These findings point to block-structured recombination rates; certain regions (called hotspots) are more prone than other regions to recombination. In this report a coalescent model incorporating hotspot or block-structured recombination is developed and investigated analytically as well as by simulation. Our main results can be summarized as follows: (1) The expected number of recombination events is much lower in a model with pure hotspot recombination than in a model with pure homogeneous recombination, (2) hotspots give rise to large variation in recombination rates along the genome as well as in the number of historical recombination events, and (3) the size of a (nonrecombining) block in the hotspot model is likely to be overestimated grossly when estimated from SNP data. The results are discussed with reference to the current debate about block-structured recombination and, in addition, the results are compared to genome-wide variation in recombination rates. A number of new analytical results about the model are derived.

THE process of recombination in humans has been intensively debated over the last years. Various recent findings suggest that the standard model assuming a flat rate of recombination along a chromosome is too crude an approximation to the actual recombination process acting on the human genome and that the standard model does not adequately explain the findings. In Daly *et al.* (2001), Jeffreys *et al.* (2001), Johnson *et al.* (2001), and Gabriel *et al.* (2002) it is argued that recombination tends to happen more often in certain regions, so-called hotspots, of a chromosome than in other regions, giving rise to long islands of nonrecombining or virtually nonrecombining genetic material.

If the above reports are true, our understanding of the recombination process as an evolutionary force must be adjusted accordingly: Modeling of recombination and interpretation of recombination patterns plays an important role in the analysis of genetic data. In this report we develop a model, the coalescent with recombination hotspots, which can be used for simulation and analysis of genetic data. Simulation of genetic data is an important tool for investigating and testing hypotheses about how genetic data have been shaped and is a useful way of gaining intuition about and insight into the consequences of evolutionary processes.

The coalescent with recombination hotspots is an extension of Kingman's (1982) coalescent and of the coalescent with recombination in various forms, the coalescent with uniform recombination rate and multilocus coalescent models (Hudson 1983; Hudson and Kaplan 1985; see Griffiths 1981 for a two-locus model). The idea is that recombination breakpoints are not chosen randomly along the chromosomes but are concentrated in certain regions of the chromosomes. One way to model this is to choose centers of recombination activity (*i.e.*, hotspots) according to some point process (*e.g.*, a Poisson process) and let recombination events happen at a rate descending from the centers. In the following a model along these lines is developed. In the next section an informal description of the model is presented followed by a mathematical treatment with comparisons to the standard model. A scheme for simulation of sequence samples and histories is described. Some familiarity with the coalescent with recombination is required.

This report is intended to be methodological, where issues of relevance to the analysis of data are addressed. The new model is compared to the coalescent model with uniform recombination rate through simulations of various summary statistics. Of special interest are the consequences of ignoring hotspot recombination and how hotspot recombination affects the genome-wide variation in recombination rates. Various issues relating to inference in the hotspot model are raised in the discussion.

A MODEL OF RECOMBINATION HOTSPOTS

Think of an entire chromosome as being represented by a line and the gene or region we are interested in as being represented by the interval (0, 1), as illustrated in Figure 1. Throughout we use "gene" in a loose sense, letting it be short for an arbitrary but fixed region in

[1]*Corresponding author:* Variagenics, 60 Hampshire St., Cambridge, MA 02139. E-mail: wiuf@birc.dk

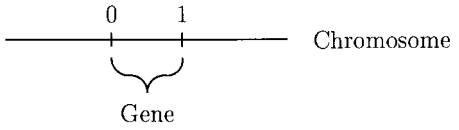

FIGURE 1.—Chromosome and gene.

the genome. Thus, a gene might be several thousand kilobases long. For a gene L nucleotides long each nucleotide takes up $1/L$ of the interval. However, for mathematical convenience we think of the gene as a continuous stretch of points.

The model we outline below is fairly general. In MATHEMATICAL EXPOSITION we restrict the model to a simpler model that still has most of the flexibility of the general model. Choose hotspots according to some point process. Perhaps the simplest and most sensible process in this connection is a Poisson process with intensity $\lambda > 0$, so that on average there are λx hotspots in a chromosomal segment of length x, and in particular there are λ hotspots on average in the gene (0, 1). In this fashion hotspots are scattered throughout the chromosome and different genes will have different numbers of hotspots, but different copies of the same chromosome will have the exact same number and the exact same locations of hotspots. In Figure 2 hotspots are labeled x_j, $j = \pm 1, 2, \ldots$. If we knew the exact locations of the hotspots, *e.g.*, from experiments, these would not have to be modeled stochastically. In the absence of such knowledge the point process reflects our prior information or expectation of how hotspots are distributed throughout the chromosome.

Recombination crossovers happen around a particular hotspot, x_j, with a rate, c_j, per generation and when a crossover occurs the breakpoint is chosen according to a distribution, $g_j(x)$, around x_j. We choose to call x_j a hotspot, though a more correct terminology might be a "center of recombination activity." Unless the distribution $g_j(x)$ is closely centered around x_j, few recombinations would be at x_j precisely. We say that recombination happens around x_j if the breakpoint is chosen from $g_j(x)$. The rates c_j could be chosen from some distribution, *e.g.*, a Γ, or be constant for all j, $c_j = c$. In the former case we talk about rate heterogeneity, and in the latter about rate homogeneity. Similarly, $g_j(x)$ might vary with j or be independent of j, $g_j(x) = g(x)$. For example, $g_j(x)$ could be normal with variance σ_j^2, or

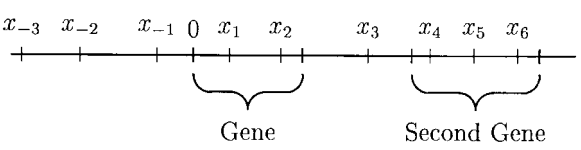

FIGURE 2.—Genes and hotspots. Each point x_j, $j = \pm 1, 2, \ldots$, represents a hotspot. Those to the left of 0 are indexed by negative integers, those to the right by positive integers.

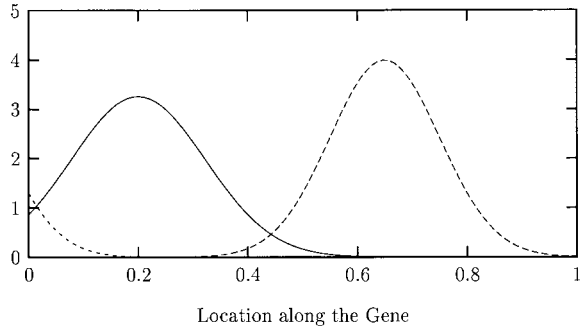

FIGURE 3.—Example of gene with two hotspots at $x_1 = 0.2$ and $x_2 = 0.65$. In addition, there is a hotspot outside the gene at $x_{-1} = -0.15$. Given that a recombination happens around x_j, $j = -1, 1, 2$, the breakpoint is chosen according to a normal density $N(0, \sigma_j^2)$ with $\sigma_j^2 = 0.01$, 0.015, and 0.01, respectively. The three curves shown (one only partly) are all normal.

$g_j(x)$ could be uniform on $(-\alpha_j, \alpha_j)$. The parameters σ_j^2 and α_j could be chosen from a set of values or from a distribution. Potentially, this results in a model with many parameters stemming from the point process, the distribution of c_j, and the specification of $g_j(x)$. Whether a hotspot, x_j, is "hot" or "cold" (as used by, *e.g.*, ROSENBERG and NORDBORG 2002) depends on the two dimensions, c_j and $g_j(x)$: c_j determines the absolute rate of recombination in the region near x_j, whereas $g_j(x)$ determines the relative rates of recombination for positions near x_j. The "hottest" hotspots are obtained with high c_j and very peaked $g_j(x)$; the "coolest" are obtained with low c_j and flat $g_j(x)$.

Two hotspots are in the example given in Figure 3, one at $x_1 = 0.2$ and the other at $x_2 = 0.65$, and $g_j(x)$, $j = 1, 2$, are normal with variances $\sigma_1^2 = 0.015$ and $\sigma_2^2 = 0.01$, respectively. Thus, most breakpoints occur near the hotspots but some might fall farther away. There is little chance that a breakpoint around x_2 falls to the left of x_1 and vice versa, but some chance that a breakpoint around x_1 falls outside the gene and in consequence the recombination event does not affect the evolution of the gene. Also, there is positive probability that a hotspot located outside the gene at $x_{-1} = -0.15$ (not shown in Figure 3) gives rise to a breakpoint that is within the gene (indicated by the dotted line at the left in Figure 3).

Since $g_j(x)$ is proportional to the probability by which recombination happens at distance x from the hotspot x_j, the sum of the curves in Figure 3 represents the overall rate of recombination in a given point (Figure 4). If $g_j(x)$ is sufficiently narrow around hotspots little overlap with other hotspots occurs, resulting in truly distinguishable peaks.

The following interpretation of the model is intuitive: A hotspot x_j can be thought of as a specific site or segment that is required for recombination to take place; however, the breakpoint itself might not be at the hotspot or fully determined by the hotspot, but just

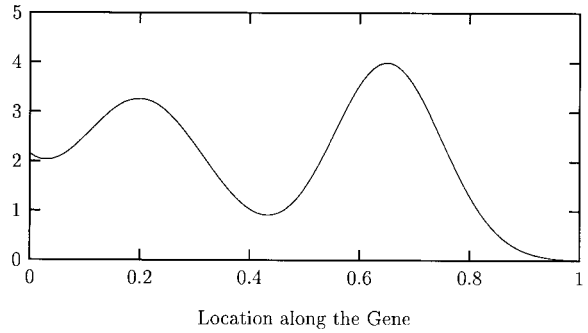

FIGURE 4.—The rate of recombination in a site z accumulated over all hotspots in Figure 3.

located somewhere randomly in the vicinity. It is worth stressing that at present published data are sparse and there is little evidence for choosing one model [*i.e.*, the point process x_j, $j = \pm 1, 2 \ldots$, the recombination rates c_j, and the breakpoint distributions $g_j(x)$] in favor of another.

MATHEMATICAL EXPOSITION

In the following we focus on establishing some results about the rate of recombination between two sites and show how the rate affects the number of segregating sites, S_n, and the number of recombination events, R_n, in a sample's history as compared to the standard model with flat recombination rate. We have simplified the general model of recombination hotspots presented in the previous section to ease presentation and computations. However, some of the results hold more generally. To be specific, we assume that $g_j(x)$ does not depend on j, *i.e.*, $g_j(x) = g(x)$ for all j, and that the distance between hotspots is Gamma distributed, $\Gamma(m, \lambda)$, $m > 0$ (hereafter referred to as the "gamma process"). If $m = 1$, then the gamma process is a Poisson process with rate λ. Allowing $m \neq 1$ introduces interference: If $m > 1$, hotspots tend to be pushed away from each other, whereas if $0 < m < 1$, they tend to cluster. Further, we assume that the distribution of rates, c_j, has expectation c. Table 1 provides an overview of the notation.

If an event happens around x_j the probability that the breakpoint is in $(z, z + dz)$ is $g(z - x_j)\,dz$, where dz denotes a small segment around z, say, of the length of a nucleotide. Summing over all hotspots x_j gives the rate $r_z dz$ by which recombination happens per generation in a particular site z. Here r_z is given by

$$r_z = \sum_{j=-\infty}^{\infty} c_j g(z - x_j). \tag{1}$$

The sum in (1) is finite, because hotspots are dropped according to the gamma process. In fact, the expectation of r_z over all possible outcomes of x_j is

$$E(r_z) = \frac{\lambda c}{m}\int_{-\infty}^{\infty} g(x)\,dx = \frac{\lambda c}{m}, \tag{2}$$

TABLE 1

List of symbols

Symbol	Explanation
λ	Scale parameter in gamma process
m	Shape parameter in gamma process
c_j	Recombination rate per generation in hotspot x_j
c	Expectation of c_j
γ_j	$4Nc_j$
γ	$4Nc$
$r_z dz$	Recombination rate per generation in site z
$r_{z_1 z_2}$	Recombination rate per generation between sites z_1 and z_2
r	Expectation of r_{01}, $r = \lambda c/m$
$\rho_{z_1 z_2}$	$4Nr_{z_1 z_2}$
ρ	Expectation of ρ_{01}, $\rho = \lambda\gamma/m$
v	Homogeneous recombination rate per gene per generation
ν	$4Nv$

which implies that (1) is finite (see the APPENDIX). Let r be $\lambda c/m$. Then, $E(r_z) = r$.

The rate, $r_{z_1 z_2}$, of recombination between any two sites, z_1 and z_2, can be found from (1) by integration,

$$r_{z_1 z_2} = \int_{z_1}^{z_2} r_z dz, \tag{3}$$

and in particular the rate for the whole gene is r_{01}. The average rate of recombination between the two sites is thus

$$E(r_{z_1 z_2}) = \int_{z_1}^{z_2} E(r_z)\,dz = r(z_2 - z_2). \tag{4}$$

Again, in particular this applies for the whole gene, $E(r_{01}) = r$. For the gamma process the average number of hotspots in the gene $(0, 1)$ is λ/m and $E(r_{01})$ is therefore the expected number of hotspots times the average rate of recombination around a hotspot.

The variances of r_z and $r_{z_1 z_2}$ are more involved and do not have simple closed expressions. In special cases they can be found, though (see the APPENDIX).

Simulation of sample histories: Consider a diploid population of size N; *i.e.*, there are $2N$ chromosomes. It follows from standard arguments (*e.g.*, HUDSON 1990) that for N large and c_j small, $c_j \approx 0$, and the time (going into the past) until a gene has been created by a recombination event is exponential with parameter $\rho_{01}/2$, $\mathrm{Exp}(\rho_{01}/2)$. Here $\rho_{z_1 z_2} = 4Nr_{z_1 z_2}$ and time is measured in units of $2N$ generations. Also define $\gamma = 4Nc$, the scaled average recombination rate per hotspot.

The location of the breakpoint z is given by the density, $h(z)$:

$$h(z) = \frac{r_z}{r_{01}} = \frac{1}{\sum_{j=-\infty}^{\infty}\int_0^1 c_j g(x - x_j)\,dx}\sum_{j=-\infty}^{\infty} c_j g(z - x_j). \tag{5}$$

The density $h(z)$ is called the breakpoint distribution

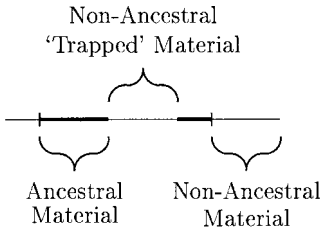

FIGURE 5.—The genetic material of a gene divided into ancestral, nonancestral, and nonancestral trapped material. Recombination affects the history of the gene if the breakpoint is in ancestral or nonancestral trapped material.

and is in general difficult to sample from given that the hotspots x_j are fixed (instead of random) and that the density $g(x)$ might have large support (a normal, for example). One uninspired way to sample $h(z)$ is to simulate the hotspots in the gene and in two extra regions surrounding the gene, thereby ignoring the rest of the chromosome. The length of the extra regions should be chosen such that a hotspot outside the two regions has little chance of producing a breakpoint inside the gene. Then for each hotspot in the gene and the two regions simulate the distribution $g(x)$ around x_j and tabulate the breakpoint distribution $h(z)$.

Simulation of sample histories can be performed in the following way. Assume that there are n genes in the sample. Let k count the number of genes that at a given time in the past carries ancestral material (see Figure 5 for an illustration).

1. Simulate c_j and x_j according to the gamma process with parameters m and λ. Calculate $h(z)$ from Equation 5 and compute ρ_{01}.
2. Simulate the next event according to the coalescent with flat recombination rate ρ_{01} and sample size k.
3. If the event is a coalescence event choose two genes at random to coalesce; otherwise choose one gene to recombine. The breakpoint, z, is chosen according to Equation 5. Update k.
4. Stop when $k = 1$.

Simulation of hotspots (step 1) is straightforward once the position of the first hotspot in the region is determined: The length between hotspots is $\Gamma(m, \lambda)$, which can be simulated using standard algorithms. The location of the first hotspot in the region can be simulated using a rejection algorithm. Details are given in the APPENDIX.

Time spent on computation in step 3 can be reduced if a look-up table for $h(z)$ is constructed. A look-up table takes the form of a dense grid of points and for each point, v, the corresponding point, z, on the gene is determined such that the probability of a recombination breakpoint between 0 and z is v, *i.e.*, P(break between 0 and z) $= \int_0^z h(x)\,dx = v$. For each recombination event a uniform variable, V, is simulated and $v = V$ is looked up in the table. The corresponding point z on the gene is the breakpoint.

It is straightforward to include uniform recombination and/or gene conversion in addition to recombination by hotspots. Uniform recombination can either be included in the breakpoint distribution $h(z)$, adjusting r_z accordingly, or be accounted for as a separate type of event. Thus, in a simulation three types of events are possible: coalescence, hotspot recombination events, and uniform recombination events. Gene conversion is best treated as a different type of event, because it often involves two breakpoints instead of one (WIUF and HEIN 2000). Both extensions seem realistic in light of how genetic data presently are conceived. Infinite-site mutation can be simulated at the same time as the genealogy, as described by previous authors (*e.g.*, HUDSON 1990 and references therein).

Number of segregating sites: Of interest is the distribution of (A) the number of segregating sites, S_n, given a particular outcome, x_j, of the gamma process and (B) the same number averaged over all possible outcomes of the gamma process. In both cases we average over all possible outcomes of c_j. B relates to the genome-wide variation, whereas A relates to the variation within a single gene. Assume the mutation rate is $\theta = 4Nu$ for the whole gene (0, 1), where u is the mutation rate per gene per generation. As shown in HUDSON (1990), S_n is Poisson-distributed $\text{Po}(\theta L_n/2)$, where L_n is the total branch length measured in $2N$ generations of the genealogy relating n sequences. The distribution of L_n depends on whether we consider a particular outcome x_j (A) or whether we average over all such outcomes (B).

The expectation of S_n under both A and B can be obtained easily, because it depends on the genealogy for a given site only (see, *e.g.*, HUDSON 1990). For A we find

$$E_x(S_n) = \theta \sum_{j=1}^{n-1} \frac{1}{j}, \tag{6}$$

where E_x denotes expectation given an outcome of x_j, $j = \pm 1, 2, \ldots$. For B we find

$$E(S_n) = E[E_x(S_n)] = \theta \sum_{j=1}^{n-1} \frac{1}{j}. \tag{7}$$

The variance is more involved because it depends on the covariance between genealogies, which in turn depends on the recombination rate between sites. Thus the variance of S_n differs under A and B. Applying the method in HUDSON (1983; see also WIUF 2000) it can be shown that the variance under A is

$$\text{Var}_x(S_n) = E_x(S_n) + \frac{\theta^2}{2} \int_0^1 \int_{z_1}^1 f_n(\rho_{z_1 z_2})\,dz_1\,dz_2, \tag{8}$$

where $f_n(x)$ denotes the covariance between the branch length of the genealogies in two sites separated by x

recombination units (Hudson and Kaplan 1985). The whole gene is ρ_{01} units. The integral differs from that of Hudson (1983) because $\rho_{z_1z_2}$ does not depend simply on the length $z_2 - z_1$ (*cf.* Equation 4). For $n = 2$, $f_n(x)$ is known,

$$f_2(x) = \frac{4(18 + x)}{18 + 13x + x^2},$$

and for $n > 2$, Kaplan and Hudson (1985) provide recursions. Under B the variance becomes

$$\mathrm{Var}(S_n) = E[\mathrm{Var}_x(S_n)] = E(S_n) + \frac{\theta^2}{2}\int_0^1\int_{z_1}^1 E[f_n(\rho_{z_1z_2})]\,dz_1dz_2 \tag{9}$$

(upon taking expectation of Equation 8; $\mathrm{Var}(E_x[S_n]) = 0$).

Both variances are bounded from above by

$$\theta\sum_{j=1}^{n-1}\frac{1}{j} + \theta^2\sum_{j=1}^{n-1}\frac{1}{j^2},$$

which is the variance of S_n for nonrecombining sequences (Watterson 1975). If $n = 2$, $f_n(x)$ is convex and using Jensen's inequality,

$$E[f_2(\rho_{z_1z_2})] \geq f_2(E[\rho_{z_1z_2}]) = f_2(\rho(z_2 - z_1)),$$

because $E(\rho_{z_1z_2}) = \rho(z_2 - z_1)$. As a consequence, $\mathrm{Var}(S_2)$ under B is larger than the variance of S_2 under a model with uniform rate ρ,

$$\theta + \frac{\theta^2}{2} \geq \mathrm{Var}(S_2) \geq \theta + \frac{\theta^2}{2}\int_0^1 (1 - x) f_2(\rho x)\,dx. \tag{10}$$

The approximation of $f_n(x)$ given in Hudson (1983) provides a similar (approximative) bound for all n. Unfortunately, the recursion provided in Kaplan and Hudson (1985) is difficult to work with and does not seem to offer a proof of (10) for general n.

Number of recombination events: The same line of argumentation as in the previous paragraph applies to the number of recombination events, R_n, in a sample's history. We show results for the expectation and variance of R_n without detailed proofs, again distinguishing between A and B.

For A we have

$$E_x(R_n) = \rho_{01}\sum_{j=1}^{n-1}\frac{1}{j}, \tag{11}$$

and thus,

$$E(R_n) = E(\rho_{01})\sum_{j=1}^{n-1}\frac{1}{j} = \rho\sum_{j=1}^{n-1}\frac{1}{j}. \tag{12}$$

The expression for the variance of R_n under A follows directly from Hudson and Kaplan (1985),

$$\mathrm{Var}_x(R_n) = E_x(R_n) + \frac{\rho_{01}^2}{2}\int_0^1 (1 - z) f_n(\rho_{01}z)\,dz. \tag{13}$$

Equations 11 and 13 follow from rescaling the gene (0, 1) so that the rate of recombination becomes uniform: The distribution of R_n given ρ_{01} is that of R_n in a model of uniform recombination rate ρ_{01}. Similarly to $\mathrm{Var}_x(S_n)$, $\mathrm{Var}_x(R_n)$ can be bounded from above:

$$\mathrm{Var}_x(R_n) \leq E_x(R_n) + \rho_{01}^2\sum_{j=1}^{n-1}\frac{1}{j^2}.$$

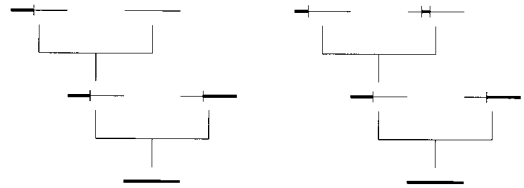

Figure 6.—A two-locus recombination model (left) compared to an infinite-site recombination model (right). The second recombination event does not affect the genealogy in the two-locus model because the second breakpoint hits the location of the first.

The variance of R_n under B cannot simply be obtained from (13) by taking expectation, because $E_x(R_n)$ varies with x_j. Instead, using the definition of the variance, we find

$$\mathrm{Var}(R_n) = E[\mathrm{Var}_x(R_n)] + \mathrm{Var}(\rho_{01})\left(\sum_{j=1}^{n-1}\frac{1}{j}\right)^2, \tag{14}$$

or

$$\mathrm{Var}(R_n) = E(R_n) + E\left[\frac{\rho_{01}^2}{2}\int_0^1 (1 - z) f_n(\rho_{01}z)\,dz\right] + \mathrm{Var}(\rho_{01})\left(\sum_{j=1}^{n-1}\frac{1}{j}\right)^2. \tag{15}$$

The bound on $\mathrm{Var}_x(R_n)$ provides a bound on $\mathrm{Var}(R_n)$,

$$\mathrm{Var}(R_n) \leq E(R_n) + E(\rho_{01}^2)\sum_{j=1}^{n-1}\frac{1}{j^2} + \mathrm{Var}(\rho_{01})\left(\sum_{j=1}^{n-1}\frac{1}{j}\right)^2.$$

In general, neither $\mathrm{Var}(\rho_{01})$ nor the second term in (15) can be calculated explicitly. For $n = 2$, the function $x^2f_n(x)/2 + x^2$ is convex and it follows that $\mathrm{Var}(R_2)$ is always larger than the variance of R_2 in a model with uniform rate ρ. The last term is almost zero if the support of $g(x)$ covers a large region. It attains its largest value if $g(x)$ has all probability mass in the hotspots. If $m = 1$ (Poisson process), then $\mathrm{Var}(\rho_{01}) = \lambda\gamma^2 = \rho^2/\lambda$ and $E(\rho_{01}^2) = \lambda(\lambda + 1)\gamma^2 = \rho^2 + \rho^2/\lambda$ (see also the appendix).

Equations 11–13 hold only if $g(x)$ is a continuous distribution; *e.g.*, if the breakpoint is exactly in x_j, then $E_x(R_n)$ and $E(R_n)$ are lower than the values given in Equations 11 and 12, respectively. In Figure 6, left, the second breakpoint hits the location of the first and, thus, does not count in R_n. In contrast, both events count in Figure 6, right. When all breakpoints are exactly in x_j, $j = \pm1, 2 \ldots$, the model is effectively a multilocus model rather than an infinite-site model. Griffiths (1991) discussed a two-locus model with recombination

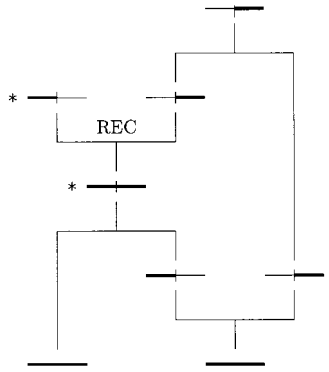

FIGURE 7.—The number of recombination events in a two-locus model with $n = 2$. Sequences marked with * are common ancestors of the sample in locus A. The recombination event marked REC does not affect the genealogy of the sample and therefore should not be counted. However, it counts in the expectations provided in GRIFFITHS (1991) and SIMONSEN and CHURCHILL (1997).

and found a recursion for the expected number of recombination events. This recursion can be solved explicitly for small n. Also SIMONSEN and CHURCHILL (1997) discussed a two-locus model and found analytic expressions for the expectation and variance of the number of recombination events for $n = 2$. Their result for the expectation agrees with that of GRIFFITHS (1991) for $n = 2$. In a sense, the number of events is overcounted in both articles, because events involving loci that already have found a most recent common ancestor are counted (Figure 7 gives an example). However, GRIFFITHS' (1991) recursion can still be applied after changing the boundary conditions (see the APPENDIX for details). Let $\varepsilon_n(y)$ be the expected number of recombination events between two loci with scaled rate y. Then

$$\varepsilon_2(y) = \frac{2y(2y + 9)}{y^2 + 13y + 18} \leq \min(y, 4), \qquad (16)$$

if $n = 2$, and

$$E_x(R_n) = \sum_{j:0<x_j<1} \varepsilon_n(\gamma_j), \qquad (17)$$

because all events that count in R_n necessarily happen between two adjacent loci. Equation 16 demonstrates that $\varepsilon_2(y)$ is lower than that in a model of uniform rate y and, more importantly, that $\varepsilon_2(y)$ does not depend linearly on the rate of recombination. A similar behavior is expected for $E_x(R_n)$ in Equation 17 or if $g(x)$ is continuous) with low variance. The expectation under B can be derived from Equation 17,

$$E(R_n) = \frac{\lambda}{m} \int_y \varepsilon_n(y) f(y)\, dy, \qquad (18)$$

where $f(y)$ denotes the density of γ_j, $j = \pm 1, 2, \ldots$. In particular, if $\gamma_j = \gamma$ is constant, $E(R_n) = (\lambda/m)\varepsilon_n(\gamma)$. It appears that $\varepsilon_2(y)$ is concave and increasing in y, and by application of Jensen's inequality,

$$E(R_2) \leq \frac{\lambda}{m} \varepsilon_2(\gamma). \qquad (19)$$

Thus, rate heterogeneity decreases the expected number of recombinations in comparison to a model with rate homogeneity. Evaluation of $E(R_n)$ in (18) can be achieved by combining numerical evaluation of the recursion for $\varepsilon_n(y)$ with Monte Carlo simulation of $f(y)$.

SIMULATION RESULTS

In this section we investigate various quantities by simulation. Particularly, we are interested in the expectation and variance of R_n and the variances of S_n and ρ_{01}. The expectation of S_n is independent of the recombination process and is given in Equation 7. The expectation of ρ_{01} is found from Equation 4. In addition, we present simulated results for HUDSON and KAPLAN's (1985) lower bound, R_M, on the number of recombination events in a sample history and MYERS and GRIFFITHS' (2003) haplotype-based bound, H_M. It is always true that $H_M \geq R_M$ and that H_M is lower than the true minimum (MYERS and GRIFFITHS 2003). Both statistics give indications to what the overall rate of recombinations might be.

The coalescent with recombination hotspots and homogeneous recombination rate was implemented in a program by one of us. A program was implemented to calculate $E(R_n)$ in Equation 18 with Gamma-distributed rates, $\gamma_j = \gamma Z_j$, $Z_j \sim \Gamma(\zeta, \zeta)$. (This is analogous to how heterogeneity in mutation rates is modeled; see, *e.g.*, YANG 1996.) Tables 2 and 3 show the mean and variance, respectively, of R_n for various combinations of the two rates: the hotspot rate ρ (varying λ, m, and γ) and the uniform rate $\nu = 4Nv$, where v is the rate per gene per generation. The density $g(x)$ was either a normal, $N(0, \sigma^2)$, or a uniform, $U(-\alpha, \alpha)$. Values of α and σ were chosen such that the respective distributions had the same variance, *i.e.*, $\alpha = \sqrt{3\sigma^2}$. Note that $\pm 2\sqrt{\sigma^2}$ creates an approximate 95% confidence interval for the normal distribution. *E.g.*, for $\sigma^2 = 5 \times 10^{-3}$, $\pm 2\sqrt{\sigma^2} = \pm 0.14$ or 14% of the whole gene. In that case $\alpha = 0.12$. Table 4 shows the variance of S_n, and Table 5 the variance of ρ_{01}; both are simulated under the same conditions as in Tables 2 and 3. The results obtained with the uniform distribution were very similar to the results obtained with the normal distribution (they deviated $< \pm 5\%$) and are thus not shown in the tables. In Tables 2–5 the sample size ($n = 20$) and the mutation rate ($\theta = 10$) are fixed; other parameter values showed similar trends. If the chosen parameter values are interpreted in the context of humans, they correspond roughly to a region of size 1000–10,000 nucleotides. This is of course not a large genomic region, but it suffices to illustrate some points.

The tables show some interesting trends. First of all, Table 2 shows that $E(R_n)$ is considerably lower in a model

TABLE 2

The expectation, $E(R_n)$, of R_n

	Homo						Het	
λ/m	0:8	4:4	8:0	0:32	16:16	32:0	0:8	4:4
1	14.10	22.96	28.38	29.65	77.99	113.5	11.94	21.77
4	20.69	25.87	28.38	56.39	91.84	113.5	18.40	24.83
16	25.38	27.54	28.38	82.78	103.5	113.5	24.06	26.95

Sample size is $n = 20$ and $\sigma^2 = 0$. Above each column is shown the values of ν and ρ as ν:ρ; Homo, rate homogeneity; Het, rate heterogeneity ($\zeta = 1$). The expectations for 8:0 and 32:0 are the theoretical values (*cf.* Equation 12), which also apply for $\sigma^2 > 0$. All other expectations are obtained from Equations 17 and 18. $E(R_n)$ depends on λ and m only through λ/m. Rate heterogeneity lowers the expectation compared to rate homogeneity.

with pure block recombination than in a model with pure homogeneous recombination. This is true irrespective of whether or not interference is assumed or whether all hotspots have the same recombination rate or the rate varies ($\zeta = 1$). The expectation $E(R_n)$ is the sum of the expectations for pure block recombination and for pure homogeneous recombination. Note that rate heterogeneity lowers the expectation compared to the case of rate homogeneity. This is in line with the observation in Equation 19 for $n = 2$. If $g(x)$ has low variance, the expectation for pure block recombination can be seen as an approximate "effective" number of recombinations, because mutations are unlikely to separate close recombination breakpoints. For $m = 1$, the probability of no hotspots within the gene is $\exp(-\lambda)$, which evaluates to 37% for $\lambda = 1$ and 2% for $\lambda = 4$. For $m = 4$, the probability of no hotspots is $\exp(-\lambda)\{1 + \frac{3}{4}\lambda + \frac{1}{4}\lambda^2 + \frac{1}{24}\lambda^3\}$, which is 20% for $\lambda = 4$ and 0.1% for $\lambda = 16$.

Table 3 shows that variation in R_n is largely induced by variation in the number of hotspots; as λ increases the rate becomes more uniform and the variance decreases, both for $m = 1$ and $m = 4$. As an example compare the three values for $\sigma^2 = 10^{-4}$ in the first column. This is further emphasized by the observation that $\mathrm{Var}(R_n)$ seems to be (slowly) decreasing in $\sigma^2 > 0$, which is as expected because the recombination rate becomes more uniform with higher σ^2. Interference decreases variation, because the variance in the number of hotspots decreases with increasing m and λ/m fixed, and the recombination rate becomes more uniform. Heterogeneity, on the other hand, increases variation, because the recombination rate becomes more variable.

The same trends in Table 3 are seen in Table 4: $\mathrm{Var}(S_n)$ decreases with increasing σ^2 and also with increasing λ, for both $m = 1$ and $m = 4$. However, note that for $\sigma^2 = 0$ the variance of S_n attains its largest value because trees for individual nucleotides are more correlated than those for $\sigma^2 > 0$. We do not see the same for R_n: If $\sigma^2 = 0$, some recombination events break

TABLE 3

The variance, $\mathrm{Var}(R_n)$, of R_n

		$\nu + \rho = 8$						$\nu + \rho = 32$:	
		$m = 1$, Homo		$m = 4$, Homo		$m = 1$, Het		$m = 1$, Homo	
λ/m	σ^2	0:8	4:4	0:8	4:4	0:8	4:4	0:32	16:16
1	0	218.7	116.3	101.3	68.99	205.6	130.7	908.3	628.3
	10^{-4}	766.5	254.4	358.3	139.1	1461	506.7	9455	3086
	5×10^{-3}	746.7	248.3	363.8	135.7	1399	388.1	10070	2883
4	0	136.0	93.15	65.31	65.14	153.3	106.8	868.4	635.5
	10^{-4}	249.6	119.0	121.5	90.49	475.4	177.6	3088	1094
	5×10^{-3}	239.2	117.1	126.2	85.60	432.5	172.9	3129	1043
16	0	98.75	75.47	58.75	64.94	100.3	79.53	624.7	459.1
	10^{-4}	117.9	88.74	83.77	78.44	169.3	99.28	1189	596.4
	5×10^{-3}	121.1	81.39	80.34	74.77	162.5	93.63	1101	579.0

Sample size is $n = 20$. Above each column is shown the values of ν and ρ as ν:ρ; Homo, rate homogeneity; Het, rate heterogeneity ($\zeta = 1$). A total of 1000 replicates were obtained for each entry. The variances for 8:0 and 32:0 (pure homogeneous recombination) are 73.94 and 420.6, respectively, obtained from simulation of 1000 replicates.

TABLE 4

The variance, Var(S_n), of S_n

		$\nu + \rho = 8$						$\nu + \rho = 32$:	
		$m = 1$, Homo		$m = 4$, Homo		$m = 1$, Het		$m = 1$, Homo	
λ/m	σ^2	0:8	4:4	0:8	4:4	0:8	4:4	0:32	16:16
1	0	163.0	117.0	162.3	117.7	167.0	112.7	140.3	74.78
	10^{-4}	153.5	117.7	136.6	111.4	172.6	114.3	141.9	71.83
	5×10^{-3}	151.8	103.8	139.4	102.5	146.6	109.0	133.1	74.83
4	0	124.2	100.9	114.5	109.9	136.4	107.3	106.7	71.33
	10^{-4}	122.1	103.7	111.4	109.9	127.8	108.1	92.86	65.90
	5×10^{-3}	120.8	101.1	96.12	104.2	118.0	110.0	90.50	64.36
16	0	110.3	105.5	105.9	104.6	107.6	96.58	74.48	66.15
	10^{-4}	111.3	106.2	102.5	106.2	104.8	104.5	70.95	65.31
	5×10^{-3}	107.7	105.5	98.8	100.1	100.7	101.1	70.98	63.96

Sample size is $n = 20$ and $\theta = 10$. The expectation of S_n is 38.48. Above each column is shown the values of ν and ρ as ν:ρ; Homo, rate homogeneity; Het, rate heterogeneity ($\zeta = 1$). A total of 1000 replicates were obtained for each entry. The variances for 8:0 and 32:0 (pure homogeneous recombination) are 95.45 and 66.21, respectively, obtained from simulation of 1000 replicates.

between the same nucleotides and might therefore not count in R_n (*cf.* Figure 6). Interference and heterogeneity affect the variance similarly to what was seen in Table 3, though less dramatically.

Table 5 summarizes the genome-wide variation, Var(ρ_{01}), in recombination rates in the hotspot model. The variation is mainly due to variation in the number of hotspots and very little to the value of σ^2. If λ is large the variance approaches 0, but it can be substantial for small values of λ. It appears that the variance is roughly inversely proportional to λ.

TABLE 5

The variance, Var(ρ_{01}), of ρ_{01}

		Homo		Hetero
λ/m	σ^2	$m = 1$	$m = 4$	$m = 1$
1	0	100.0	40.78	200.0
	10^{-4}	103.3	39.33	193.6
	5×10^{-3}	96.45	33.86	190.5
4	0	25.00	7.23	50.00
	10^{-4}	24.23	6.95	49.74
	5×10^{-3}	22.96	6.08	49.94
16	0	6.25	1.62	12.50
	10^{-4}	6.22	1.56	12.38
	5×10^{-3}	5.55	1.38	11.75

The variance of ρ_{01} for $\rho = E(\rho_{01}) = 10$. Homo, rate homogeneity; Hetero, rate heterogeneity ($\zeta = 1$). The variance for other values of ρ is obtained by multiplying the value in the table by $\rho^2/10^2$. Note that adding homogeneous recombination does not change Var(R_n), because the homogeneous rate is constant. The variances for $\sigma^2 = 0$ are obtained theoretically (see the APPENDIX). A total of 1000 replicates were obtained for each entry. It appears that the values for $m = 4$ are roughly a factor of 4 smaller than those for $m = 1$, Homo, and that the variance is inversely proportional to λ for large λ.

Table 6 shows summary statistics of R_M and H_M. The inferred minimum number of recombinations is lower in the hotspot model than in the uniform model: We tend to underestimate the amount of historical recombinations more in the hotspot model than in the uniform model. This is the case even if $\sigma^2 > 0$ (results not shown), because recombination events around a hotspot are detected by R_M or H_M only if there is an accumulation of mutations in the region around the hotspot. The actual number of recombination events is in general much higher (*cf.* Table 2).

DISCUSSION

We have developed an extension of the coalescent with recombination that in a simple way accounts for heterogeneity in recombination rate and hotspots. Rate heterogeneity can be modeled in ways other than the one proposed here: Basically, what is required is a relationship, stochastic or deterministic, between physical and genetic distance. The standard coalescent model of uniform rate suggests a linear relationship between the two distances. This results in a one-parameter model. In contrast the model proposed here has a stochastic relationship between the two distances, in the sense that the rate is different for each realization of a gene's history. The model can be summarized as having four main parameters: an intensity, λ, that controls the number of hotspots; a measure of interference, m; a recombination rate, γ, per hotspot; and a parameter that regulates the size, σ^2, of the hotspot. Two of these, σ^2 (or more correctly, $\sqrt{\sigma^2}$) and λ, are scaled in the

TABLE 6

Summary statistics of R_M and H_M

		R_M						H_M:	
		$n = 20, \theta = 10$		$n = 50, \theta = 10$		$n = 20, \theta = 40$		$n = 20, \theta = 10$	
ν	ρ	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
8	0	2.33	1.70	2.84	1.80	4.31	3.67	3.16	3.31
4	4	2.10	1.46	2.47	1.64	3.81	2.85	2.92	3.14
0	8	1.55	1.10	1.72	1.24	2.28	1.84	2.37	3.34
32	0	4.97	3.34	5.81	3.51	10.4	7.14	8.20	8.28
16	16	4.10	2.41	4.99	3.00	8.37	6.12	7.07	7.48
0	32	2.16	1.39	2.38	1.66	3.14	2.55	4.31	6.81

Sample size $n = 20, 50$; mutation rate $\theta = 10, 40$. Hotspots are simulated with $\sigma^2 = 0$, $\lambda = 4$, and $m = 1$. The expectation of S_n is 38.48 for $\theta = 10$ and $n = 20$, 153.9 for $\theta = 40$ and $n = 20$, and 44.79 for $\theta = 10$ and $n = 50$. A total of 1000 replicates were obtained for each entry.

length of the gene, γ is scaled in the effective population size, and m has no scale.

Fisher *et al.* (1947) introduced the gamma process in the context of chromatid interference. Their model is called the χ^2 model. The χ^2 model is not conceptually identical to the model presented in this article: In our model the gamma process determines the location of hotspots. These locations are the same for all individuals and are potential breakpoints for recombination. In contrast, in the χ^2 model the gamma process determines the location of chromatid crossovers, actual recombination breakpoints. These vary from individual to individual, both in number and in location.

We showed how various summary statistics are affected by hotspot recombination compared to uniform recombination. One important message is that in the hotspot model we tend to underestimate the number of recombination events more in a sample history than in a model of homogeneous recombination. From an inference point of view this is extremely unsatisfactory: There are more parameters to estimate in the hotspot model than in the uniform model and reliable estimates might therefore be difficult to achieve. For one thing, it is well known that estimators of the recombination rate are often biased downward (Wall 2000), suggesting that these estimators are even more likely to be biased downward in the hotspot model. As a further issue, the extra parameters in the hotspot model make inference computationally intractable. Even in the standard model, maximum-likelihood estimation is computationally nontrivial (Fearnhead and Donnelly 2001), and many approaches rely on summary statistics and/or simplifications of the likelihood (Wall 2000; Fearnhead and Donnelly 2002, and references therein).

Statistically, it is not known whether the parameters can be estimated consistently. Fearnhead (personal communication) shows that in certain cases it is possible to estimate the flat recombination rate consistently as the length of the genomic region increases. Similarly, it might be expected that λ, m, γ, and σ^2 can be estimated consistently. Fearnhead's (personal communication) proof does not directly apply in the present situation because he works with the standard one-parameter coalescent model and the proof makes explicit use of this. As a final comment along these lines, if data are analyzed under the standard model alone, a systematic downward bias is expected; this was clearly demonstrated in Table 2.

We calculated summary statistics for the recombination rate from 22 genes spread throughout the human genome (from Table 1 in Nachman 2001). We found an average rate of 2.08 cM/Mb with a variance of 1.63. If we assume the effective population size is $N = 10^4$ and a gene of $\sim 10^4$ nucleotides, then estimates of the expectation and variance of ρ_{01} are given by $E(\rho_{01}) = 8.32$ and $\mathrm{Var}(\rho_{01}) = 26.08$, respectively. Compared with the values in Table 5, these point to a fairly high level of variation in recombination rate. For example, the variance ($26.08 \times 10^2/8.32^2 = 37.7$) is consistent with some interference and homogeneity in rates (*e.g.*, $m = 4$, $\lambda = 4$) or with no interference and heterogeneity in rates (*e.g.*, $m = 1$, $\lambda = 5.2$, assuming an inverse proportional relationship between $\mathrm{Var}(\rho_{01})$ and λ; see Table 5).

Table 6 also brings a message to researchers spending effort on inferring haplotype maps of human chromosomes. Hotspots (or block end and start points) are inferred from single-nucleotide polymorphism (SNP) patterns, but as shown in Table 6 these patterns lead to gross underestimation of the true number of recombination events. The statistic R_M infers at most one recombination between any two SNPs: $\hat{\lambda} = R_M$ is thus an estimate of λ (in general, R_M is an estimate of λ/m) and $\hat{B} = 1/R_M$ an estimate of block size. If $\nu = 0$ and $\rho = 8$, then $\hat{B} = 0.65$ for $\theta = 10$, and $\hat{B} = 0.43$ for $\theta = 40$. If $\nu = 0$ and $\rho = 32$, then $\hat{B} = 0.46$ for $\theta = 10$ and $\hat{B} = 0.32$ for $\theta = 40$ (using the simulated expected values of R_M). However, the expected block size is $1/\lambda = 0.25$

($\lambda = 4$) and in consequence $\hat{B}$ overestimates block size by at least 80% for $\theta = 10$ and by at least 30% for $\theta = 40$. For $n = 50$ the block size becomes more accurately estimated than for $n = 20$, though still overestimated by at least 60% ($\theta = 10$). For pure homogeneous recombination the expected number of topology-changing recombinations is 1.49ν for $n = 20$ and 2.37ν for $n = 50$ (Hudson and Kaplan 1985). Thus R_M and H_M underestimate the amount of recombination grossly. It is noteworthy that on average there are >150 segregating sites if $\theta = 40$. This is unlikely to be the case for most genes. The statistic H_M might infer more than one recombination event between two SNPs and is therefore less useful as an estimator of block size. In a recent article (Zhang *et al.* 2002), block size is even more severely overestimated. In some cases their method fails to take obvious recombination hotspots into account. Consider, for example, the four haplotypes

0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 1 1 1 1 1
1 1 1 1 1 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1

(10 SNPs are shown here, but one could add more). Zhang *et al.*'s (2002) method finds one block, but obviously there are two since all four gametes are present. This example is not artificial at all: SNPs with this pattern are shown in Daly *et al.* (2001). Further research will hopefully shed more light on these issues as they will continue to be of relevance and interest to researchers in genetics for years to come. Particularly, we have in mind the efforts, initiated by the National Human Genome Research Institute (http://www.genome.gov), to accomplish a haplotype map of the human genome.

We thank L. Subrahmanyan for helpful and useful discussions. We are grateful to two anonymous reviewers who provided many fruitful comments and suggestions. The program used for simulation can be obtained from http://www.uvigo.es/dposada.

LITERATURE CITED

Daly, M., J. D. Rioux, S. F. Scaffner, T. J. Hudson and E. S. Lander, 2001 High-resolution haplotype structure in the human genome. Nat. Genet. **29:** 229–232.

Ethier, S., and R. C. Griffiths, 1990 On the two-locus sampling distribution. J. Math. Biol. **29:** 131–159.

Fearnhead, P., and P. Donnelly, 2001 Estimating recombination rates from population genetic data. Genetics **159:** 1299–1318.

Fearnhead, P., and P. Donnelly, 2002 Approximate likelihood methods for estimating local recombination rate. J. R. Stat. Soc. B **64:** 657–680.

Fisher, R. A., M. F. Lyon and A. R. G. Owen, 1947 The sex chromosome in the house mouse. Heredity **1:** 335–365.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy *et al.*, 2002 The structure of haplotype blocks in the human genome. Science **296:** 2225–2229.

Griffiths, R. C., 1981 Neutral two-locus multiple alleles model with recombination. Theor. Popul. Biol. **19:** 169–186.

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

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

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

Jeffreys, A. J., L. Kauppi and R. Neumann, 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat. Genet. **29:** 217–222.

Johnson, G. C. L., L. Esposito, B. J. Baratt, A. N. Smith, J. Heward *et al.*, 2001 Haplotype tagging for the identification of common disease genes. Nat. Genet. **29:** 233–237.

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

Karlin, S., and H. M. Taylor, 1975 *A First Course in Stochastic Processes.* Academic Press, New York.

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

McPeek, M. S., and T. P. Speed, 1995 Modeling interference in genetic recombination. Genetics **139:** 1031–1044.

Myers, S. R., and R. C. Griffiths, 2003 Bounds on the minimum number of recombination events in a sample history. Genetics **163:** 375–394.

Nachman, M. W., 2001 Single nucleotide polymorphisms and recombination rate in humans. Trends Genet. **17:** 481–485.

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. **3:** 380–390.

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

Wall, J. D., 2000 A comparison of estimators of the population recombination rate. Mol. Biol. Evol. **17:** 156–163.

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

Wiuf, C., 2000 A coalescence approach to gene conversion. Theor. Popul. Biol. **57:** 357–367.

Wiuf, C., and J. Hein, 2000 The coalescent with gene conversion. Genetics **155:** 451–462.

Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analysis. Trends Ecol. Evol. **11:** 367–372.

Zhang, K., M. Deng, T. Chen, M. S. Waterman and F. Sun, 2002 A dynamic programming algorithm for haplotype block partition. Proc. Natl. Acad. Sci. USA **99:** 7335–7339.

Communicating editor: J. Hey

APPENDIX

Expectation of r_z: Upon taking expectation of r_z, Equation 1 can be written

$$E(r_z) = E(c_j)\int_x g(z - x)\sum_j \delta_j(x)\,dx,$$

where $E(c_j) = c$, and $\delta_j(x)$ is the density of the location of the jth hotspot. The sum $\Sigma_j\,\delta_j(x)$ is independent of x and is known as the intensity of the process. For the gamma process the intensity is λ/m (*e.g.*, McPeek and Speed 1995). Thus,

$$E(r_z) = \frac{\lambda c}{m}\int_x g(z - x)\,dx = \frac{\lambda c}{m}.$$

Variance of $r_{z_1z_2}$ and $\rho_{z_1z_2}$: If $g(x)$ puts all probability

mass in the hotspot, the variance of $r_{z_1z_2}$ can be found. Note that in this particular case

$$r_{z_1z_2} = \sum_{j:z_1<x_j<z_2} c_j,$$

so that

$$r^2_{z_1z_2} = \sum_j c_j^2 + 2\sum_{i<j} c_i c_j,$$

and

$$E(r^2_{z_1z_2}) = \frac{\lambda}{m} E(c_j^2) + 2c^2 \int_0^1 \int_u^1 P(\text{hotspots in } u \text{ and } v)\, du dv$$

$$= \frac{\lambda}{m} E(c_j^2) + 2\frac{c^2\lambda}{m} \int_0^1 (1-u) \sum_{k=1}^{\infty} \eta_k(u)\, du.$$

The latter follows from McPeek and Speed (1995), but is more generally treated in Karlin and Taylor (1975). Here $\eta_k(u)$ is

$$\eta_k(u) = \frac{1}{\Gamma(mk)} \lambda^{mk} u^{mk-1} e^{-\lambda u}.$$

If $m = 1$, then $\Sigma_k \eta_k(u) = \lambda$, and

$$E(r^2_{z_1z_2}) = \lambda E(c_j^2) + c^2\lambda^2.$$

In particular, if $c_j = c$, then $E(r^2_{z_1z_2}) = \lambda(\lambda + 1)c^2 = r(r + 1/\lambda)$ and $\mathrm{Var}(r_{z_1z_2}) = \lambda c^2 = r^2/\lambda$. If $c_j = cY_j$, $Y_j \sim \Gamma(\zeta, \zeta)$, then $\mathrm{Var}(r_{z_1z_2}) = \lambda c^2(1 + 1/\zeta) = r^2(1 + 1/\zeta)/\lambda$. If $m = 2$, then $\Sigma_k \eta_k(u) = \lambda(1 - e^{-2\lambda u})/2$ and

$$E(r^2_{z_1z_2}) = \frac{\lambda}{2} E(c_j^2) + \frac{c^2}{4}[\lambda(\lambda - 1) + \frac{1}{2}(1 - e^{-2\lambda})]. \quad \text{(A1)}$$

If $m = 4$, then $\Sigma_k \eta_k(u) = \lambda(1 - e^{-2\lambda u})/4 - \lambda e^{-\lambda u}\sin(\lambda u)/2$ and

$$E(r^2_{z_1z_2}) = \frac{\lambda}{4} E(c_j^2) + \frac{1}{4}C(\lambda) - \frac{c^2}{8}[\lambda - 1 + e^{-\lambda}\cos(\lambda)],$$

where $C(\lambda)$ is the second term in Equation A1. The variance is easily obtained from these equations. By replacing r with ρ and c with γ, the variance of $\rho_{z_1z_2}$ is obtained.

Simulation of the first hotspot in a gene: McPeek and Speed (1995) state the distribution of the position of the first hotspot in a gene,

$$\delta_1(x) = \frac{\lambda^{m+1}}{\Gamma(m+1)} \int_u^{\infty} u^{m-1} e^{-\lambda u} du, \quad \text{(A2)}$$

and provide a rejection algorithm for simulating from $\delta_1(x)$. The density of x_{-1} is $\delta_{-1}(x) = \delta_1(x)$. If m is an integer, a value from $\delta_1(x)$ can be realized by simulating a Gamma variable, $\Gamma(k, \lambda)$, where k is chosen randomly among $1, \ldots, m$. If m is noninteger, let m_0 be the first integer $> m$, and simulate a value, x, from $\delta_1(x)$ with m_0 and λ. Accept x with probability

$$\frac{(m_0 - 1)!}{\lambda^{m_0 - m}\Gamma(m)} \frac{\int_x^{\infty} u^{m-1} e^{-\lambda u} du}{\int_x^{\infty} u^{m_0 - 1} e^{-\lambda u} du}. \quad \text{(A3)}$$

Equation A3 can be evaluated only numerically.

Recursion for $\varepsilon_n(y)$: Griffiths (1991) provides the following recursion for calculation of $\varepsilon_n(y)$ for given values of n and y. Let a, b, and $c \geq 0$, and let $n = a + b + c$. The number c is the number of genes where both loci are ancestral to the sample, a is the number where locus 1 is ancestral, and b is the number where locus 2 is ancestral. The recursion is

$$q(a, b, c; y) = \frac{cy}{n(n-1) + cy}[1 + q(a+1, b+1, c-1; y)]$$

$$+ \frac{2ab}{n(n-1) + cy} q(a-1, b-1, c+1; y)$$

$$+ \frac{a(a + 2c - 1)}{n(n-1) + cy} q(a-1, b, c; y)$$

$$+ \frac{b(b + 2c - 1)}{n(n-1) + cy} q(a, b-1, c; y)$$

$$+ \frac{c(c-1)}{n(n-1) + cy} q(a, b, c-1; y), \quad \text{(A4)}$$

with boundary conditions $q(0, b, 1; y) = q(1, b, 0; y) = q(a, 0, 1; y) = q(a, 1, 0; y) = 0$. Then $\varepsilon_n(y) = q(0, 0, n; y)$. [Griffiths (1991) uses boundary conditions $q(0, 0, 1; y) = q(0, 1, 0; y) = q(1, 0, 0; y) = q(1, 1, 0; y)$.] Ethier and Griffiths (1990) solve (A4) using a tridiagonal scheme.