Modeling Interference in Genetic Recombination

by

Mary Sara McPeek, Terence P. Speed

Department of Statistics, University of California,

Berkeley, California 94720

Running Head: Modeling Interference

Corresponding author:

Mary Sara McPeek

address: The University of Chicago Department of Statistics,

5734 University Ave., Chicago, IL 60637

phone: (312) 702-7554

FAX: (312) 702-9810

ABSTRACT

In analyzing genetic linkage data, it is common to assume that the locations of crossovers along a chromosome follow a Poisson process, whereas it has long been known that this assumption does not fit the data. In many organisms, it appears that the presence of a crossover inhibits the formation of another nearby, a phenomenon known as "interference." We discuss several point process models for recombination which incorporate position interference, but assume no chromatid interference. Using stochastic simulation, we are able to fit the models to a multilocus Drosophila dataset by the method of maximum likelihood. We find that some biologically-inspired point process models incorporating one or two additional parameters provide a dramatically better fit to the data than the usual "no-interference" Poisson model.

The phenomenon of interference in genetic recombination was noticed very early this century by *Drosophila melanogaster* geneticists in Thomas Hunt Morgan's lab (STURTEVANT 1915; MULLER 1916). They found that simultaneous recombination in two or three nearby chromosomal intervals occurred much less often than would be expected under independence, and that the effect appeared to decrease with distance. At present, the biological nature of genetic interference is still not well-understood, nor has it been adequately modeled, mathematically. Virtually all multilocus linkage analyses use the assumption of no interference. While this assumption may give consistent results for locus ordering (SPEED, MCPEEK and EVANS 1992) and estimation of recombination fractions in some cases, it clearly does not fit the data. It is natural to ask if such an analysis could be improved by the use of a reasonable interference model. What is required is a biologically plausible point process model for crossovers along a chromosome, which should be fit to data. The Drosophila data set of MORGAN, BRIDGES and SCHULTZ (1935) is ideal for this purpose.

Previous attempts to fit a crossover point process to the MORGAN, BRIDGES and SCHULTZ (1935) Drosophila dataset (COBBS 1978; RISCH and LANGE 1983; PASCOE and MORTON 1987; GOLDGAR and FAIN 1988; FOSS et al. 1993) have been severely limited by the difficulty of calculating the likelihood of the data under all but the simplest models. Using a Monte Carlo method, we are able to fit a wide range of point process models to the data.

BACKGROUND

In diploid eukaryotes, crossing over takes place during the pachytene phase of meiosis, when the two homologous versions of any particular chromosome have each been duplicated, and all four resulting strands or chromatids are lined up together, forming a very tight bundle. We model crossovers as being points located along this bundle, and each crossover involves exactly two of the four chromatids. We assume that the two chromatids involved in any particular crossover are nonsister chromatids, that is, the two chromatids cannot be the two copies of a single homologous chromosome. After crossing over has occurred, the four resulting chromatids are mixtures of the original parental types. In Drosophila, for each meiosis, we will have information on only one of these four resulting chromatids. If that chromatid was involved in an odd number of crossovers between two loci, a recombination is said to have taken place between the two loci.

It is important to keep in mind that crossing over takes place among four chromatids. In that case, the two aspects relevant to recombination are (i) the distribution of crossovers along the bundle of four chromatids and (ii) which pair of nonsister chromatids is involved in each crossover. The concept of interference is usually divided into two parts, corresponding to aspects (i) and (ii). First, we say that there is position or chiasma interference if the number and location of crossovers in a given region are not independent of the numbers and locations of crossovers in disjoint regions.

Second, we say there is chromatid interference if it is not the case that each pair of nonsister chromatids is equally likely to be involved in a crossover, independent of which were involved in other crossovers. There is little consistent evidence of chromatid interference (H. ZHAO, M. S. MCPEEK, and T. P. SPEED, in press).

We assume no chromatid interference and try to find a better-fitting point process model than the Poisson for the occurrence of crossovers along the chromosome. In the case of no chromatid interference, the point process on the single chromatid can be obtained from the point process on the bundle of four by independently thinning (i.e. deleting) each point with probability $\frac{1}{2}$. This is because the given chromatid has chance $\frac{1}{2}$ of being involved in a particular crossover, independent of involvement in any others. As a result of assuming no chromatid interference, the chance of recombination across an interval increases monotonically as the interval is enlarged, with an upper bound of $\frac{1}{2}$.

The traditional measure of interference is the coincidence (MULLER 1916; STURTEVANT 1915), which is expressed as a ratio. The numerator is the chance of simultaneous recombination across both of two disjoint intervals on the chromosome. The denominator is the product of the marginal probabilities of recombination across the intervals.

$$C = \frac{r_{11}}{(r_{10} + r_{11})(r_{01} + r_{11})} ,$$

where C is the coincidence and r_{ij} is the chance of i recombinations across the first interval and j recombinations across the second interval. If there were no position interference, the coincidence would equal one. Observed coincidences tend to be near zero for small, closely-linked intervals, increasing to one for more distant intervals.

The recombination point processes considered here are stationary in terms of some distance metric, although this distance metric is generally not equivalent to physical distance. This point will be discussed further below. For each of the point process models considered here, there is more than one such choice of a distance metric which will make the process stationary; these metrics are the same up to a multiplicative constant. The genetic distance associated with a chromosome interval is defined to be the expected number of crossovers occurring on a single chromatid within that interval. The metrics in which the models discussed here are stationary are all constant multiples of genetic distance. We shall choose one for notational convenience and call it the stationary metric. Among other things, stationarity of the model means that coincidence will be a function of the distances across each of the two intervals considered and the distance between the intervals (in terms of the stationary metric), but not their actual locations. This also implies that the intensity, $\mu = \lim_{h \to 0} \; h^{-1} Pr(\text{at least 1 point in } [t, t+h))$, where h is measured in terms of the stationary metric, is the same for all t. Furthermore, for the models considered here, the conditional intensity can be defined by $\mu(z) = \lim_{\delta \to 0} \lim_{\epsilon \to 0} \; \epsilon^{-1} Pr(\text{at least 1 point in } [t+z, t+z+\epsilon) \mid \text{at least 1 point in } [t, t+\delta))$, where ϵ is measured in terms of the stationary metric. Intuitively, this is the intensity at $t+z$ conditional on a point at t, and by stationarity, this depends only on z. When the conditional intensity exists, the coincidence function can be defined as the coincidence between two intervals, in the limit when the stationary widths of the intervals are allowed to go to zero, as a function of the stationary distance between the two intervals. The coincidence function is given by the ratio of the conditional intensity to the unconditional intensity

$$C(z) = \frac{\mu(z)}{\mu}.$$

Formulae of this type appear in FOSS et al. (1993) and LANDE and STAHL (1993). Note that coincidence is not a complete description of interference, but measures interference only between pairs of intervals.

Multilocus Linkage Analysis: When one has a panel of genetic markers, one may be interested in making a linkage map, that is, in ordering the markers and calculating genetic distances between them. Alternatively, one may be interested in using linkage analysis to locate a new marker or gene of interest on a previously determined marker map. In both cases, the most informative kind of analysis is multilocus linkage analysis, in which one considers recombination patterns among all the marker loci simultaneously.

The Drosophila dataset of MORGAN, BRIDGES and SCHULTZ (1935) consists of counts of recombination events among nine marker loci on the X chromosome. The nine loci under consideration are *scute, echinus, crossveinless, cut, vermilion, sable, forked, carnation,* and *bobbed,* and they correspond to observed fly phenotypes. The actual positions of the loci on the X chromosome are assumed to be fixed but unknown, and counts of recombination events are made based on the observed physical characteristics of the fly, which are associated with alleles at the nine loci. In *Drosophila melanogaster*, recombination occurs only in females. Consider the X chromosome inherited by a fly from its mother. This X chromosome will be some mixture of the two maternal grandparental X chromosomes, due to crossing over between them. For each of the nine loci, it can be determined whether the offspring fly inherited its maternal grandmother or maternal grandfather's allele at that locus. Thus, there are 2^9 possible observed outcomes. However, each outcome has a complementary outcome which is considered equivalent, in terms of recombination, namely the one in which all the grandmaternal and grandpaternal alleles are switched. Therefore, there are 2^8 possible recombination outcomes, and the number of times each occurs out of 16,136 meioses is recorded in the MORGAN, BRIDGES and SCHULTZ dataset.

If the nine loci are ordered, we let A_j denote the interval between loci j and j+1. We let

$x = (x_1, ..., x_8)$, $x_j = 0$ or 1, j=1,...8, denote the event of a recombination in each interval A_j for which $x_j = 1$, and no recombination in each interval A_j for which $x_j = 0$. Then each of the 2^8 possible recombination outcomes would correspond to one of the 2^8 possible x's.

In what follows, we assume the order of the loci to be known. If the order were not known, the procedures described below could be repeated for each of a small number of candidate orders, and the estimated order would be the one whose maximized likelihood was highest. We assume that each event x occurs with some fixed probability p_x. These events correspond to what is observed in the data. For each x, we wish to calculate its probability p_x under each of the point process models we consider, in order to fit the models to the data by maximum likelihood.

Recall that event x occurs when a given chromatid is involved in an odd number of crossovers in each of the intervals A_j for which $x_j = 1$ and an even number of crossovers in each of the intervals A_j for which $x_j = 0$. A set of related events will be denoted $y = (y_1, ..., y_8)$, $y_j = 0$ or 1, j=1,...,8. y is the event that, on the bundle of four chromatids, at least one crossover occurs in each of the intervals A_j for which $y_j = 1$ and no crossovers occur in each of the intervals A_j for which $y_j = 0$. Note that y is an event occurring on the bundle of four chromatids, while x refers to just one of the four chromatids. We let q_y denote the probability of the event y. The assumption of no chromatid interference gives a correspondence between these two sets of probabilities (WEINSTEIN 1936), namely

$$p_x = \sum_{y:\ 1 \geq y \geq x} \frac{1}{2^{y \cdot 1}} q_y \qquad \text{for all } x$$

and inverting,

$$q_y = \quad 2^{y \cdot 1} \times \sum_{x:\ 1 \geq x \geq y} (-1)^{(x-y) \cdot 1} p_x \qquad \text{for all } y,$$

where, for example, $y \cdot 1 = \sum_{j=1}^{8} y_j$, and $1 \geq y \geq x$ means $1 \geq y_j \geq x_j$ for all j. Thus, to get the recombination probabilities under different point process models, assuming no chromatid interference, it would suffice to calculate probabilities for the simpler events, y, corresponding to zero or nonzero crossovers in the locus intervals, on the bundle of four chromatids. Once we can calculate the q_y's in terms of the unknown parameters and from these, the p_x's, we write down the log-likelihood as

$$\text{constant} + \sum_x a_x log(p_x),$$

where $x = (x_1, ..., x_8)$, $x_j = 0$ or 1, a_x is the number of times the event x occurs in the data and where the constant is irrelevant since it does not involve the unknown parameters. We would then maximize the likelihood over the unknown parameters.

MODELS

It is sometimes assumed that the problem of modeling interference is equivalent to that of formulating a map function, i.e. a function from $[0,\infty)$ to $[0,\frac{1}{2}]$ which maps expected number of crossovers to chance of a recombination. This is not the case. Firstly, a map function is not a full-fledged model. By itself, a map function does not provide probabilities for multilocus recombination events when more than three loci are involved, and so it cannot be fit to data such as that described above. In an attempt to remedy this problem, LIBERMAN and KARLIN (1984) introduced multilocus map functions, which are defined as above, but with the additional property that the map function relationship between expected number of crossovers and chance of a recombination should hold on unions of disjoint intervals, as well as on intervals. Unfortunately, these multilocus map functions correspond to an extremely limited class of models, under the assumption of no chromatid interference. EVANS, MCPEEK and SPEED (1992) showed that under this assumption, the only models which have multilocus map functions are the count-location models described by KARLIN and LIBERMAN (1979) and RISCH and LANGE (1979). These models have the undesirable property that the coincidence function is constant, i.e. the level of interference, as measured by coincidence, does not vary at all with the genetic distance between the intervals considered. In actual data, coincidence seems to be close to zero for near intervals and close to one for more distant intervals (see Fig. 2). For this reason, the count-location model does not appear well-suited to modeling interference. It is, however, easy to fit, and has been previously fit to the Drosophila data of MORGAN, BRIDGES and SCHULTZ (RISCH and LANGE, 1983). We include it here for comparison. The model will be described in more detail below.

Poisson Model: COX and ISHAM (1980) provides an introduction to point processes. The simplest crossover point process model we consider is the Poisson process. This model was proposed for recombination by HALDANE (1919) and continues to be virtually the only model used in linkage analysis. The model allows no interference at all, i.e. crossing-over in disjoint intervals is independent, or the presence of one crossover does not alter the chance of others occurring nearby. Thus, the coincidence function has the constant value 1. In the general formulation of the Poisson model, we let μ_t denote the intensity of the process at physical location t along the bundle of four chromatids, i.e.

$$\mu_t = \lim_{h \to 0} \frac{P(\text{at least 1 point in } (t, t+h])}{h}.$$

We shall not consider cases in which this limit does not exist, nor in which μ_t is not integrable. Then the chance of no crossovers in a given interval [a,b) is $\exp(-\int_a^b \mu_t dt)$, or $e^{-\mu(b-a)}$ in the homogeneous case where $\mu_t = \mu$ for all t. To get the probability of a given event y, i.e. the event of at least one crossover in each interval for which $y_i = 1$ and no crossovers in each interval for which $y_i = 0$, we multiply the appropriate probabilities for each interval, using independence.

If we let $l_1, ..., l_9$ denote the physical locations of the nine marker loci, then from recombination data, the estimable quantities in the Poisson model are the genetic distances (i.e. expected numbers of crossovers on a single chromatid) between the markers, namely,

$$d_1 = \frac{1}{2}\int_{l_1}^{l_2} \mu_t dt, \quad d_2 = \frac{1}{2}\int_{l_2}^{l_3} \mu_t dt, \quad ..., \quad d_8 = \frac{1}{2}\int_{l_8}^{l_9} \mu_t dt.$$

(The factor of $\frac{1}{2}$ occurs because each crossover on the bundle of four chromatids is assumed to involve a given chromatid with chance $\frac{1}{2}$.) Thus, from the data, one cannot infer anything about possible inhomogeneity of the Poisson intensity, because this cannot be separated from the unknown locations of the marker loci. Similarly, one cannot estimate the actual locations of the markers nor the physical distances between them. Associating the interval [0,1] on the real line with the chromosomal segment between the outermost loci under consideration, and letting μ_t and $0 = l_1 < ... < l_9 = 1$ be as before, let $\mu = \int_0^1 \mu_t dt$ and $M(z) = \mu^{-1} \int_0^z \mu_t dt$, for all $z \in [0,1]$. Then $M : [0,1] \to [0,1]$ is monotone non-decreasing and onto (i.e. for every point $w \in [0,1]$, there is some $z \in [0,1]$ such that $w = M(z)$), hence continuous. M can be thought of as a continuous cumulative distribution function (cdf) on [0,1]. If we transform each point of the Poisson process by M, then the resulting process is homogeneous Poisson with intensity μ. The transformed homogeneous process on [0,1] with marker locations $M(l_1) < ... < M(l_9)$ and the original inhomogeneous process on [0,1] with marker locations $l_1 < ... < l_9$ would both produce recombination data with the same distribution. Thus, without loss of generality, we may consider only homogeneous Poisson processes, that is, we let $\mu_t = \mu$ for all t. The estimable parameters of the model are the genetic distances, $d_1, ..., d_8$, between the markers and in terms of these, the probability of the event y is

$$\prod_{i=1}^{8} e^{-2d_i y_i}(1 - e^{-2d_i})^{(1-y_i)}.$$

Note that μ can be written in terms of this parametrization: $\mu = 2\sum_{i=1}^{8} d_i$.

Gamma Model: The distances between crossovers in the homogeneous Poisson model are independent exponential random variables, or equivalently, gamma with shape parameter 1. One way to generalize the Poisson model would be to consider renewal processes with general gamma interarrivals, thus adding an extra parameter to the model. On the real line, the stationary renewal process with gamma interarrivals can be formulated as follows: given a point at location w, the density of the distance to the next point to the right is

$$f_I(z) = \frac{1}{\Gamma(\gamma)}\mu^\gamma z^{\gamma-1}e^{-\mu z},$$

where I stands for "interarrival", $\Gamma(\gamma) = \int_0^\infty \mu^\gamma s^{\gamma-1}e^{-\mu s}ds$, and this is independent of the occurrence of any points to the left of w. The distribution of the next point to the left of w is the same; there is no directionality. The distribution of the distance to the first point to the right (or equivalently, left) of w, when it is not assumed that a point has occurred at w, has density

$$f_{1st}(z) = \frac{\mu}{\Gamma(\gamma+1)}\int_z^\infty \mu^\gamma s^{\gamma-1}e^{-\mu s}ds.$$

The intensity of the process is then $\frac{\mu}{\gamma}$. Letting $\gamma = 1$, we would get the Poisson model. The coincidence function for the gamma model is

$$C(z) = \frac{\gamma}{\mu}\sum_{n=1}^{\infty} f_I^n(z),$$

where $f_I^n(z) = \Gamma(\gamma n)^{-1}\mu^{\gamma n}z^{\gamma n-1}e^{-\mu z}$.

We associate the interval [0,1] with the bundle of four chromatids, and we consider the above process restricted to [0,1]. Then the chance of at least one crossover occurring on the bundle of four chromatids is $\alpha = \int_0^1 f_{1st}(z)dz$, and given that at least one crossover occurs, the density of the distance from the 0 end of the bundle to the first crossover is $\alpha^{-1}f_{1st}(z)$. Given a crossover at location $w \in [0,1]$, the chance that another crossover occurs between w and 1 is $\beta(w) = \int_0^{1-w} f_I(z)dz$.

Given a crossover at location w, and given that at least one crossover occurs between w and 1, the density of the distance from w to the next crossover between w and 1 is $\beta(w)^{-1}f_I(z)$. Note that all distances are in terms of a metric whose relationship with physical distance is unknown, but which is a constant multiple of genetic distance.

We consider the gamma model on the bundle of four chromatids, and the points are then independently thinned, each with chance $\frac{1}{2}$, (that is, each point has chance $\frac{1}{2}$ to be deleted, independently of which other points are deleted) to get the crossover point process on a single chromatid. We do not restrict consideration to the particular case of gamma with integer shape parameter, but the use of the gamma renewal process with integer shape parameter to model crossover occurrence has a long history. FISHER, LYON and OWEN (1947) proposed that the crossover point process along a single chromatid be viewed as a renewal process with interarrival density $f(z) = \frac{1}{2}\pi\text{sech}(\frac{1}{2}\pi z)\tanh(\frac{1}{2}\pi z)$. OWEN (1949;1950) found that a gamma interarrival density with shape parameter 2 and scale parameter 2 was mathematically tractable and closely approximated the renewal process proposed by FISHER, LYON and OWEN. CARTER and ROBERTSON (1952) used the gamma with shape parameter 2 as a four-strand model and applied various models of chromatid interference to get the chance of recombination across an interval for a single chromatid. PAYNE (1956) considered the gamma with integer shape parameter as a two-strand model. He compared coincidence curves for the gammas with shapes 2 and 3 to data. COBBS (1978) considered the gamma with integer shape parameter as a four-strand model, assuming no chromatid interference. He fit the model to Neurospora and Drosophila data by comparing the observed distribution of the number of crossovers in a segment (ignoring those which could not be observed) to the number predicted by the model. STAM (1979) considered various mathematical aspects of the same model, and FOSS et al. (1993) also considered this model, comparing the coincidence curves from the model to coincidence curves for Drosophila and Neurospora data. In this paper, we calculate approximate probabilities of the different possible multilocus recombination events under the model, and we fit the model to the data by maximum likelihood. We are able to compare observed to expected frequencies of recombination events. This subsumes both the comparisons of distributions of number of events and of coincidence curves.

The gamma interarrival process with integer shape parameter has been used in the literature so often largely because of relative mathematical tractability. However, FOSS et al. (1993) propose it in order to explain certain empirical observations concerning recombination and gene conversion (a non-reciprocal exchange of genetic material between homologous chromosomes). First, gene conversion is associated with a high frequency of recombination of flanking markers (MORTIMER and FOGEL 1974). Second, gene conversions seem to occur independently in disjoint intervals, but gene conversions accompanied by recombination do not, rather, they appear to inhibit each other (MORTIMER and FOGEL). The model proposed by FOSS et al. (see also STAHL (1979), MORTIMER and FOGEL (1974)) is that initial pre-crossover events occur along the chromosome according to a Poisson process. Each such event results in a gene conversion and may or may not result in a crossover as well. Their model holds that every $(m+1)$st initial event results in a crossover. If the first event has chance $\frac{1}{m+1}$ to form a crossover, then the model is equivalent to the stationary gamma interarrival model with shape parameter $m+1$.

We note that, as in the case of the Poisson model, if we fit the stationary gamma renewal process, we are also allowing for the possibility that the true physical process may be nonstationary. The class of models covered by the analysis consists of those for which a monotone non-decreasing, onto transformation M: [0,1] $\rightarrow$ [0,1] (i.e. a continuous, cdf on [0,1]) exists which maps the model to a stationary model. The models corresponding to almost-everywhere differentiable M's are those which can be formulated as follows: given a point at w, the density of the distance to the first point to the right of w is

$$f_I^w(z) = \frac{\mu_z}{\Gamma(\gamma)}\left(\int_w^z \mu_s ds\right)^{\gamma-1} e^{-\int_w^z \mu_t dt},$$

where $\mu_t = \mu \times \frac{dM(t)}{dt}$. The density of the distance to the first point to the right of a given location w, when it is not assumed that a point has occurred at w, is

$$f_{1st}^w(z) = \int_{-\infty}^w \frac{\mu_t}{\gamma} f_I^t(z)dt.$$

The intensity of the process at location t is $\gamma^{-1}\mu_t$. Note that the coincidence function is identical to that for the stationary case.

The estimable parameters are the shape parameter γ and the genetic distances $d_1, ..., d_8$. Note that μ can be written, in terms of this parametrization, as $\mu = 2\gamma \sum_{i=1}^8 d_i$. It has not proved possible to write down an explicit expression for the likelihood of the data under the gamma model, except in the case of integer shape parameter (H. ZHAO, T. P. SPEED and M. S. MCPEEK, in press). Instead, the likelihood has been simulated, as described below.

Hard-Core Model: Another generalization of the Poisson model would be to have the points follow a Poisson process, but with no two points allowed to be closer than a certain fixed distance h. This is known as a hard-core model (see e.g. STOYAN, KENDALL, and MECKE 1987). On the real line, this is just a stationary renewal process with interarrivals distributed as exponential plus a constant. That is, given a point at location w, the distance to the next point to the right (equivalently, left) has density

$$f_I(z) = \begin{cases} \mu e^{-\mu(z-h)} & \text{for } z > h \\ 0 & \text{otherwise} \end{cases} .$$

The distribution of the distance to the first point to the right (equivalently, left) of w, when it is not assumed that a point has occurred at w, has density

$$f_{1st}(z) = \begin{cases} \frac{\mu}{1+\mu h} & \text{if } z \leq h \\ \frac{\mu e^{-\mu(z-h)}}{1+\mu h} & \text{if } z > h \end{cases} .$$

The intensity of the process is $\frac{\mu}{1+\mu h}$. Letting $h = 0$, we would get the Poisson model. The coincidence function for the hard-core model is

$$C(z) = \frac{\mu}{1+\mu h} \sum_{n=1}^{\infty} f_I^n(z),$$

where

$$f_I^n(z) = \begin{cases} \Gamma(n)^{-1} \mu^n (z - nh)^{n-1} e^{-\mu(z-nh)} & \text{if } z \geq nh \\ 0 & \text{otherwise} \end{cases} .$$

Again, we associate the interval [0,1] with the bundle of four chromatids. The restriction of this stationary renewal process to the interval [0,1] works exactly as in the gamma case. As before, this hard-core model is equivalent, in terms of the data, to any model resulting from a transformation of the interval by a continuous cdf M. The models corresponding to almost-everywhere differentiable M's are those which can be formulated as follows: given a point at w, the density of the distance to the first point to the right of w is

$$f_I^w(z) = \begin{cases} \mu_z e^{-\int_{w+h_w}^{z} \mu_s ds} & \text{if } z > w + h_w \\ 0 & \text{otherwise} \end{cases} ,$$

where $\mu_t = \mu \times \frac{dM(t)}{dt}$, and $h_w > 0$ is any solution of $\int_w^{w+h_w} \mu_s ds = B$,

where $B = h\mu$. The density of the distance to the first point to the right of a given location w, when it is not assumed that a point has occurred at w is

$$f_{1st}^w(z) = \begin{cases} \frac{\mu_z}{1+B} & \text{if } z < w + h_w \\ \mu_z e^{B - \int_0^z \mu_s ds} & \text{if } z \geq w + h_w \end{cases} .$$

The intensity of the process at location t is $\frac{\mu_t}{1+B}$.

The estimable parameters are the hard-core genetic distance, $h_g = \frac{1}{2}B/(1+B)$, and the genetic distances between the points, $d_1, ..., d_8$. Note that μ can be expressed in terms of this parametrization as $\mu = (1 - 2h_g)^{-1} \sum_{i=1}^{8} d_i$. As the number of loci increases, the likelihood quickly becomes very complicated, and no tractable general form has been found.

King-Mortimer Model: In one version of a model suggested by KING and MORTIMER (1990), points are put down on a segment of the real line according to a Poisson process with parameter μ, just as in the original stationary Poisson model described above. Starting from time 0, each point independently waits an exponential amount of time with parameter λ before starting polymer growth, i.e. the density of the time to polymer growth is $f(t) = \lambda e^{-\lambda t}$. When a polymer starts to grow from a point, it grows in both directions at a constant rate g. If a polymer from one point hits another point which has not yet started to grow, the latter point is deleted. At first glance, this process appears to be nearly identical to the one-dimensional version of the Johnson-Mehl model (JOHNSON and MEHL 1939, but MEIJERING 1953 is more readable) for random tessellations. In that model, points are born over time and immediately start growing polymers (called "crystals" by JOHNSON and MEHL 1939) at rate g in both directions. In the Johnson-Mehl model, the rate of birth in an unpolymerized interval $[w, w + h_1)$ during an interval of time

$[t, t+h_2)$ is $\rho h_1 h_2 + o(h_1 h_2)$. That is, in any infinitesimal interval of time $[t, t+h_2)$, it is a Poisson process, with parameter ρh_2, on the unpolymerized part of the line. In any small unpolymerized interval $[w, w+h_1)$ along the line, one waits an exponential amount of time, with parameter ρh_1 for a point to appear and start growing. However, in the King-Mortimer model, the chance that a polymer starts to grow in a small unpolymerized interval $[w, w+h_1)$ during an interval of time $[t, t+h_2)$ can be shown to be $\mu \lambda e^{-\lambda t} h_1 h_2 + o(h_1 h_2)$, i.e. it varies over time, going monotonically to zero.

In the King-Mortimer model, there is an initial Poisson distribution of points, then the polymerizing process takes place over time, and then there is some resulting final distribution of points. Intuitively, it is clear that the final distribution of points would not be affected if time were slowed down or speeded up by some constant multiple c, as long as the initial Poisson distribution of points remained the same. This would be equivalent to changing λ to $\frac{1}{c}\lambda$ and changing g to cg. Therefore, it is not surprising that the final distribution of points is determined by μ and $\frac{g}{\lambda}$ only. Thus, without loss of generality, we may take $\lambda = 1$.

KING and MORTIMER (1990) choose parameter values for which their simulations appear to resemble observed data, but they do not actually fit the model. They consider the process as defined on a line segment. In that case points are more likely to occur near the ends of the segment, because there are no points outside the segment to interfere with them. This is also true in the case of the hard-core model defined on a segment. We instead consider the process as defined on the real line, and then restricted to [0,1]. In that case, the points near the ends of the segment behave as if other points lying outside the segment could interfere with them, and the model is stationary. As before, a non-stationary model resulting from a transformation of the interval [0,1] by a continuous cdf M is indistinguishable, in terms of the data, from the stationary model. This non-stationarity is of a different type from that introduced by considering the process only on an interval, as KING and MORTIMER (1990) have done. The nonstationary models with almost-everywhere differentiable M's covered by this analysis are those in which the initial Poisson distribution has inhomogeneous parameter $\mu_t = \mu\frac{dM(t)}{dt}$, and the growth rate of a polymer which has reached position t is $g_t = g\frac{dM(t)}{dt}$. We note that the stationary process is not a renewal process; given a point at $w \in [0,1]$, the distribution of the distance to the next point to the right of w is not independent of the occurrence of other points to the left of w.

In order to specify the stationary King-Mortimer model on [0,1], it is convenient to determine the distribution of the time when a polymer growing from somewhere outside [0,1] would reach that interval. In order to do this, we ignore, for the moment, points inside [0,1] which could potentially polymerize and prevent any outside polymers from ever reaching [0,1]. To calculate the distribution of the time when a polymer from $(1, \infty)$ reaches 1, we let X_i, for each $i > 1$, be the distance between the ith and (i-1)st initial points to the right of 1. If X_1 is the distance from 1 to the first initial point to the right of 1, then the X_i are independent exponential(μ). Let Z_i be the waiting time for polymer growth of the ith initial point to the right of 1. The Z_i are independent exponential(1). Let $S_i = g^{-1}\sum_{j=1}^{i} X_j$. Then the chance that no polymer arrives at 1 from the right before time t is

$$P(S_1 > t) + \sum_{i=1}^{\infty} P(S_i \le t < S_{i+1},\ S_j + Z_j > t,\ \text{for all } j \le i)\,.$$

The ith term of the sum is found to be

$$e^{-\mu g t}\frac{1}{i!}\,[\mu g(1 - e^{-t})]^i,$$

and so the required probability is

$$e^{\mu g(1-t-e^{-t})}.$$

One minus this quantity is the cdf of the time that a polymer first arrives from the right (equivalently, left) at any given point. This cdf is useful for simulating the process. Substituting $2g$ in for g gives the cdf of the time that a polymer first arrives at any given point from either the right or the left, ignoring the possibility that the point itself grows first and prevents the polymer's arrival. From this, we can calculate the intensity of the point process, which is just the intensity, μ, of the initial Poisson process multiplied by the chance that a given initial point starts to grow before another polymer arrives. This is

$$\mu \int_0^\infty e^{-t} e^{2\mu g(1-t-e^{-t})} dt = \mu[(\frac{e}{a})^a \,\Gamma(a,a) - \frac{1}{a}],$$

where $a = 2\mu g$ and $\Gamma(x,y) = \int_0^y t^{x-1}e^{-t}dt$, known as the incomplete gamma function. The estimable parameters under this version of the King-Mortimer model are the genetic distances, $d_1, ..., d_8$, and the expected time for a polymer to traverse the distance between two adjacent points in the initial Poisson process, as a multiple of expected waiting time for a polymer to start growing: $T = (\mu g)^{-1} = 2a^{-1}$. Note that μ can be expressed in terms of this parametrization:

$$\mu = 2 \sum_{i=1}^{8} d_i/[(\frac{eT}{2})^{2/T}\,\Gamma(2/T, 2/T) - T/2].$$

No closed-form expression is known for the likelihood of recombination data under this model, nor for the coincidence function. Both of these are obtained here by simulation.

K-M II Model: The second King-Mortimer model (KING and MORTIMER 1990), which shall be called here K-M II, is the same as the previous model except that a parameter is added for termination of polymer growth, once it has started. That is, under the set-up of the previous model, each polymer now grows for an independent exponential($1/\nu$) amount of time, or until it hits another polymer, whichever happens first. The possibility of early termination of a polymer has the effect of allowing interference to operate over a smaller range than before. Interference can then be made more intense, yet more localized than in the ordinary King-Mortimer model. For this model, there is no known expression for the intensity, the likelihood of recombination data, nor for the coincidence function. The parameters estimated are the genetic distances, $d_1, ..., d_8$, the expected time for a polymer to traverse the distance between two adjacent points in the initial Poisson configuration (given that growth does not terminate up to that time), T, and the expected time to terminate polymer growth, ν, where each of these times is expressed as a multiple of the expected waiting time for a polymer to start growing.

Count-Location Model: In the count-location model (KARLIN and LIBERMAN 1979, called the "generalized no interference model" by RISCH and LANGE 1979), the number of crossovers is chosen according to some count distribution c, where c is given by $c_0, c_1, ...$, and c_i = P(i crossovers). Given the number of crossovers, their locations are independent and identically distributed (i.i.d.) along the bundle of chromatids, according to a location distribution ν which does not vary with the count. The count-location model can be thought of as a modification of the Poisson process model, in which the number of crossovers occurring is no longer Poisson, but their locations are again i.i.d. as in the Poisson model. In the count-location model, the coincidence function is constant:

$$C(z) = \sum_{i=2}^{\infty} i(i-1)c_i / (\sum_{j=1}^{\infty} jc_j)^2.$$

The version of the count-location model considered here allows no more than three crossovers to occur simultaneously, i.e. $\sum_{i=1}^{3} c_i = 1$ and $c_i = 0$ for $i > 3$. The location distribution ν may be taken to be uniform without loss of generality since the class of models covered by this analysis will include all those with location distributions having a continuous cdf. The intensity of the process is $\sum_{i=0}^{3} ic_i$. The model can be parametrized in terms of the genetic distances, $d_1, ..., d_8$, and c_0 and

c_1. The likelihood can be calculated explicitly in terms of these parameters.

To fit any of these models to recombination data, it is necessary to be able to calculate q_y for any y, that is, the probability of any combination of zero and non-zero crossover counts in the eight marker intervals, as described above. This is trivial for the Poisson model and the count-location model. For the other models, this involves integrating the density over all possible realizations compatible with y, assuming that the density is known or can be computed. So far, these integrals appear virtually impossible to do exactly, except in the special case of gamma with integer shape parameter. As a result, they have been calculated here by Monte Carlo methods. For the Poisson, gamma, count-location, and K-M II models, the recombination probabilities were computed for events among all nine loci, while for the other models, only a subset of five of the nine loci were used.

For a given model, the probability q_y depends not only on the model parameters, but also on the relative locations of the loci. For each choice of model parameters, a large number n of realizations of the point process are simulated. Using these simulations, the desired probability q_y can be estimated, for any set of locus locations, by the observed frequency of the event y. From the q's, the estimates of the probabilities of the observed recombination events can be computed, and from these, the estimated likelihood of the observed data is computed. Recombination events in which three or more recombinations took place were so rare in this data that their probabilities are very difficult to estimate. These rare events were pooled in this analysis. Holding the point process parameters fixed, the likelihood can then be numerically maximized over the locus locations by the Nelder-Mead search algorithm (NELDER and MEAD 1965). This maximization is done using the same set of n simulated realizations. For each choice of model parameters, the process can be repeated, generating a new set of n realizations and maximizing the estimated likelihood over the locus locations. The Nelder-Mead search algorithm can then be used to maximize over the model parameters. This is an extension of a method described in DIGGLE and GRATTON (1984). The number of realizations simulated, n, was taken to be 160,000 initially and was increased to 960,000 as a maximum was approached.

There are standard algorithms for simulating from exponential and gamma distributions (see RIPLEY 1987). In addition to these, we must be able to simulate from the first-point distributions of the gamma and the hard-core models, and also, in the King-Mortimer model, the distribution of the first time that a polymer from outside the $[0, 1]$ interval first hits 0 from below (or equivalently, hits 1 from above).

In the gamma case, if the shape parameter is an integer, say γ, then a random variable from the first-point distribution can be realized by simulating the sum of k independent exponential random variables, where k is chosen uniformly on 1,...,γ. For non-integer γ's, one can do rejection sampling using the next highest integer (see RIPLEY 1987 for a discussion of rejection sampling). For the hard-core model with hard-core distance h and Poisson parameter μ, the first point will be uniform on $[0, h]$ with probability $h/(\frac{1}{\mu} + h)$ and exponentially distributed on $[h, \infty)$ with probability $\frac{1}{\mu}/(\frac{1}{\mu} + h)$.

In order to simulate from the King-Mortimer model on an interval, one must be able to simulate the distribution of the first time that a polymer from outside the interval reaches the interval from the left, or identically, from the right. The distribution, which was derived in the last section, can be simulated by rejection sampling using exponential(μ) for $\mu g \leq 1$ and exponential($\frac{1}{g}$) for $\mu g \geq 1$. For the K-M II model, even this distribution is not known, but is simulated by looking at the process over a very large interval and then restricting to a small interval, where edge effects may be assumed to be negligible.

RESULTS

SPEED, MCPEEK and EVANS (1992) found that as a result of assuming no chromatid interference, certain constraints must be satisfied by the probabilities of recombination events. Namely,

$$0 \leq \sum_{y \geq x} (-1)^{(y-x)\cdot 1} p_y \qquad \text{for all } x.$$

Furthermore, for any set of p's which satisfies these constraints, there is a no-chromatid-interference model which is compatible with them. It is interesting to note that for the five loci considered, the observed frequencies of the recombination events satisfy all eight nontrivial constraints of the ones required above. Only a single observation prevents this from being true for the 37 nontrivial constraints on all nine loci. Of course, this is not proof that the assumption of no chromatid interference is correct, but simply that the data are not incompatible with this assumption.

Table 1 shows Pearson chi-square statistics for the five-locus models and Table 4 shows them for the nine-locus models. In all cases, the p-values are smaller than .01, implying significant misfit of all models. However, with the addition of a single parameter, the chi-square statistic in the five-locus case is brought down from 773 (Poisson model) to 53 (gamma model), and in the nine-locus case, from 1672 (Poisson model) to 107 (gamma model). This represents a tremendous improvement. In the five-locus case, the best-fitting model is the K-M II model, which adds two parameters to the Poisson model and has a chi-square statistic of 17, compared to 773 for the Poisson model. The K-M II model is outperformed by the gamma model in the nine-locus case. One interpretation is that K-M II does a reasonably good job of modeling interference between nearby markers, but that it is not flexible enough to simultaneously model both close and medium-range interference, as required in the nine-locus dataset.

Tables 2 and 3 contain fitted parameter values, with estimated standard deviations, for the five-locus models, and Tables 4 and 5 show them for the nine-locus models. The results shown for the Poisson are from a calculation rather than from simulation, but the simulated results are virtually identical. In the case of the gamma model, the likelihood can actually be calculated in the case of integer shape parameter (H. ZHAO, T. P. SPEED and M. S. MCPEEK, in press). In the nine-locus case, using this likelihood calculation, the best fit is obtained using shape parameter 5, which agrees with the results given here for the nine-locus case. This would correspond to four gene conversion events between each crossover in the framework described in FOSS et al. (1993). The count-location model has previously been fit to the MORGAN, BRIDGES and SCHULTZ (1935) dataset by RISCH and LANGE (1983), using a calculated, rather than simulated likelihood. In their calculation, they did not pool events in which three or more recombinations occurred. Thus, they maximized a slightly different likelihood than the one maximized here. Still, the results are very close to those obtained here. Note that the estimates of genetic distances and their standard deviations were fairly similar for all models. The Poisson model gives very similar estimates to the other models for small genetic distances (< 10 centiMorgans), but overestimates larger genetic distances relative to the other models.

The standard deviations were estimated by simply taking numerical derivatives of the recombination probabilities (p_x's in the notation given above) with respect to the estimated parameters, and using these to calculate the Fisher information matrix. A better approach, not taken here, would be to smooth the likelihood surface before calculating derivatives.

The Pearson chi-square statistic measures the overall goodness-of-fit of each model. We consider two additional criteria which do not indicate goodness-of-fit as directly as the chi-square does, but which are easy to interpret and which may be useful for understanding the performance of the models. First, we consider the number of crossovers predicted by the models. Figure 1 shows that the five-locus Poisson model predicts too few single crossover events and too many of all other types of events. This shows up in Table 1 as an underestimate of the number of single recombination events among the five loci and and overestimate of all other types of recombination events. Figure 3 shows that in the nine-locus case, the Poisson model predicts too few single and double crossover events and too many of all other types. The gamma, count-location and K-M II models all do well at matching the distribution of the number of crossovers in the data, although the K-M II model does not do nearly as well with nine loci and with five.

The other aspect of the models which we shall consider is the coincidence function which is discussed above. Figure 2 shows coincidence vs. genetic distance for the five-locus models, and Figure 4 shows it for the nine-locus models. Recall that coincidence is a measure of the degree to which one crossover inhibits the formation of another nearby. For the Poisson model, coincidence is 1, reflecting no interference. The coincidence curves for the models have been calculated (or simulated, in the case of the King-Mortimer and K-M II models) using infinitesimal intervals. The coincidence curves of the gamma model and the King-Mortimer model give the best approximation to the data in the five-locus case, while the gamma and K-M II models do best in the nine-locus case.

The strange shape of the coincidence curve for the hard-core model can be explained as follows: in the hard-core model, no points are allowed within a fixed distance of the crossover point which is assumed to lie at 0 (for the purpose of making the coincidence curve). Immediately after that fixed distance has been surpassed, one is waiting an exponential amount of time for the next point, whereas in an interval chosen at random the intensity is less than exponential because there is a chance of another crossover already having occurred nearby. This explains the spike in the coincidence curve for the hard-core model, and the hump in the coincidence curve for the gamma model can be explained similarly: as the distance from the given crossover gets close to the mode of the gamma interarrival distribution, one is more likely to have another crossover than one would be to have a crossover in a given interval chosen at random.

It should be noted that the coincidence curve for the count-location model is constant, even if the distribution of points is i.i.d but non-uniform. The reason is that the x-axis represents genetic, not physical distance, that is, it represents the expected number of crossovers, and so is automatically rescaled to be uniform. Note also the anomaly that the K-M II model fits the data better (according to chi-square) in the five-locus case than in the nine-locus case, but the coincidence curve resembles the data much more closely in the nine-locus case.

DISCUSSION

Although none of the models fit the data, several of the models were able to capture certain aspects of the data. The best overall model of the ones considered was the renewal model with gamma interarrivals. It gave reasonable estimates of the numbers of crossovers and matched the interference pattern in the data as measured by the coincidence. In terms of the chi-square statistic, the gamma model represents a huge improvement over the Poisson model. Furthermore, in the case of integer shape parameter, the gamma model may prove tractable enough to use in mapping applications.

This work represents a first attempt to systematically fit a range of point process models to recombination data. A suitable model, such as the gamma model, may be of great use in assessing the impact of interference on conventional linkage analyses and genetic mapping. It may even be possible to apply such a model in genetic mapping, thereby making more efficient use of the data.

We are grateful to B. D. Ripley for his advice on the Monte Carlo method and to F. W. Stahl for helpful discussions. This work was supported by National Science Foundation grants DMS-880237 and DMS-9113527.

LITERATURE CITED

CARTER, T. C. and A. ROBERTSON, 1952 A mathematical treatment of genetical recombination using a four-strand model. Proc. Roy. Soc. B **139**: 410-426.

COBBS, G., 1978 Renewal process approach to the theory of genetic linkage: case of no chromatid interference. Genetics **89**: 563-581.

COX, D.R. and V. ISHAM, 1980 *Point Processes.* Chapman and Hall, London and New York.

DIGGLE, P. J. and R. J. GRATTON, 1984 Monte Carlo methods of inference for implicit statistical models. JRSS B **46:** 193-227.

EVANS, S. N., M. S. MCPEEK, T. P. SPEED, 1992 A characterisation of crossover models that possess map functions. Theor. Pop. Biol. **43:** 80-90.

FISHER, R. A., M. F. LYON and A. R. G. OWEN, 1947 The sex chromosome in the house mouse. Heredity **1**: 335-365.

FOSS, E., R. LANDE, F. W. STAHL, C. M. STEINBERG, 1993 Chiasma interference as a function of genetic distance. Genetics **133:** 681-691.

GOLDGAR, D. E. and P. R. FAIN, 1988 Models of multilocus recombination: nonrandomness in chiasma number and crossover positions. Am. J. Hum. Genet. **43:** 38-45.

HALDANE, J. B. S., 1919 The combination of linkage values, and the calculation of distances between the loci of linked factors. J. Genetics **8:** 299-309.

JOHNSON, W. A. and R. F. MEHL, 1939 Reaction kinetics in processes of nucleation and growth. Trans. A.I.M.M.E. **135:** 416-458.

KARLIN, S. and U. LIBERMAN, 1979 A natural class of multilocus recombination processes and related measures of crossover interference. Adv. Appl. Prob. **11:** 479-501.

LANDE, R. and F. W. STAHL, 1993 Chiasma interference and the distribution of exchanges in *Drosophila melanogaster*. Cold Spring Harbor Symposia on Quantitative Biology **58:** 543-552.

LIBERMAN, U. and S. KARLIN, 1984 Theoretical models of genetic map functions. Theor. Pop. Biol. **25:** 331-346.

KING, J. S. and R. K. MORTIMER, 1990 A polymerization model of chiasma interference and corresponding computer simulation. Genetics **126:** 1127-1138.

MEIJERING, J. L., 1953 Interface area, edge length, and number of vertices in crystal aggregates with random nucleation. Philips Res. Rep. **8:** 270-290.

MORGAN, T. H., C. B. BRIDGES, J. SCHULTZ, 1935 Constitution of the germinal material in relation to heredity. Carnegie Institution of Washington **34:** 284-291.

MORTIMER, R. K. and S. FOGEL, 1974 Genetical interference and gene conversion, pp. 263-275 in *Mechanisms in Recombination*, edited by R. F GRELL. Plenum, New York.

MULLER, H. J., 1916 The mechanism of crossing-over. The American Naturalist **50:** 193-221,284-305,350-366,421-434.

NELDER, J. A. and R. MEAD, 1965 A simplex method for function minimisation. Comp. J. **7:** 303-313.

OWEN, A. R. G., 1949 The theory of genetical recombination, I. Long-chromosome arms. Proc. Roy. Soc. B **136:** 67-94.

OWEN, A. R. G., 1950 The theory of genetical recombination. Adv. Genet. **3**: 117-157.

PASCOE, L. and N. E. MORTON, 1987 The use of map functions in multipoint mapping. Am. J. Hum. Genet. **40:** 174-183.

PAYNE, L. C., 1956 The theory of genetical recombination: a general formulation for a certain class of intercept length distributions appropriate to the discussion of multiple linkage. Proc. Roy. Soc. B **144:** 528-544.

RIPLEY, B. D. R., 1987 *Stochastic Simulation.* John Wiley and Sons, New York.

RISCH, N. and K. LANGE, 1979 An alternative model of recombination and interference. Ann. Hum. Genet. Lond. **43:** 61-70.

RISCH, N. and K. LANGE, 1983 Statistical analysis of multilocus recombination. Biometrics **39:** 949-963.

SPEED, T. P., M. S. MCPEEK, S. N. EVANS, 1992 Robustness of the no-interference model for ordering genetic markers. Proc. Natl. Acad. Sci. USA **89:** 3103-3106.

STAHL, F. W., 1979 *Genetic Recombination. Thinking About it in Phage and Fungi.* pp. 158, 304-306, W. H. Freeman, San Francisco.

STAM, P., 1979 Interference in genetic crossing over and chromosome mapping. Genetics **92:** 573-594.

STOYAN, D., W. S. KENDALL, J. MECKE, 1987 *Stochastic Geometry and its Applications,* Wiley, New York.

STURTEVANT, A. H., 1915 The behavior of the chromosomes as studied through linkage. Zeit. f. ind. Abst. u. Vererb. **13:** 234-287.

WEINSTEIN, A., 1936 The theory of multiple-strand crossing over. Genetics **21:** 155-199.

Table 1: Observed and expected counts under the fitted models for five loci, with chi-square statistics

Legend: Triple and quadruple recombination events are quite rare. These events were pooled in the analysis because accurate probabilities for the events could not be computed by simulation.

recomb. event	observed data	expected Poisson	expected gamma	expected hard-cor	expected King-Mor	expected count-loc	expected K-M II
0000	10431	11014.1	10497.1	10837.0	10783.6	10434.0	10443.1
1000	771	596.6	738.9	682.7	686.1	777.6	767.4
0100	1579	1246.9	1537.6	1360.6	1390.0	1556.1	1560.4
0010	1221	931.3	1214.0	1058.7	1064.2	1184.7	1210.7
0001	1994	1664.1	1979.7	1685.2	1811.7	2036.4	2018.0
1100	4	67.5	1.5	13.5	25.2	16.0	7.3
1010	7	50.4	12.0	69.3	35.8	11.8	12.4
0110	4	105.4	4.0	35.1	43.1	23.4	13.2
1001	46	90.1	73.0	106.3	85.2	20.2	32.9
0101	53	188.4	68.0	208.1	138.5	42.1	48.6
0011	25	140.7	10.1	71.7	64.8	31.5	20.8
1110 1101 1011 0111 1111	1	40.3	.1	7.9	7.8	2.3	1.3
chi-sq df		773 8	51 7	420 7	267 7	67 6	17 6

Table 2: Fitted point process parameters with estimated standard deviations for five loci

model	parameter(s)
gamma	$\gamma = 6.41$ (.280)
hard-core	$h_g = .116$ (.004)
King-Mortimer I	$T = .038$ (.005)
count-location	$c_0 = .311$ (.005), $c_1 = .641$ (.007)
	(redundant parameters: $c_2 = .0448$ (.005), $c_3 = .0032$ (.002))
K-M II	$T = 1.01 \times 10^{-4}$ (2.44×10^{-5}), $\nu = .244$ (.187)

Table 3: Fitted genetic distances between adjacent markers with estimated standard deviations for five loci

genetic distances	Poisson	gamma	hard-core	K-M I	ct-location	K-M II
(sc)-(ec)	.054 (.002)	.051 (.002)	.055 (.002)	.052 (.002)	.052 (.002)	.050 (.003)
(ec)-(cv)	.114 (.003)	.100 (.003)	.100 (.003)	.101 (.002)	.104 (.002)	.102 (.005)
(cv)-(ct)	.085 (.003)	.077 (.003)	.077 (.002)	.076 (.002)	.079 (.002)	.079 (.003)
(ct)-(v)	.152 (.004)	.132 (.004)	.129 (.003)	.135 (.003)	.136 (.003)	.132 (.007)

Table 4: Point Process Parameters and Chi-Square Statistics for 9-locus fit

model	parameter(s)	chi-sq	d.f.
Poisson	none	1672	30
gamma	γ =4.94 (.124)	107	29
count-location	c_0 = .057 (.008), c_1 = .434 (.017)	889	28
	redundant parameters: c_2 = .434 (.017), c_3 = .075 (.008)		
K-M II	$T = 1.84 \times 10^{-4}$ (7.30×10^{-6}), $\nu = 1.08$ (.337)	294	28

Table 5: Genetic distances for nine-locus fit

genetic distances	Poisson	gamma	count-loc	K-M II
(sc)-(ec)	.054 (.002)	.053 (.002)	.053 (.002)	.046 (.002)
(ec)-(cv)	.114 (.003)	.098 (.002)	.107 (.003)	.096 (.004)
(cv)-(ct)	.085 (.003)	.075 (.002)	.082 (.002)	.078 (.003)
(ct)-(v)	.152 (.004)	.133 (.002)	.141 (.003)	.141 (.005)
(v)-(s)	.089 (.003)	.085 (.002)	.086 (.002)	.088 (.004)
(s)-(f)	.184 (.004)	.157 (.003)	.171 (.004)	.163 (.007)
(f)-(car)	.081 (.002)	.075 (.003)	.078 (.002)	.071 (.003)
(car)-(bb)	.046 (.002)	.044 (.002)	.044 (.002)	.039 (.002)

Figure Legends

Figure 1:

Distribution of the number of crossovers among the first five loci of the MORGAN, BRIDGES and SCHULTZ (1935) dataset, for each of the models under consideration, relative to the data distribution. Figure 1a shows the chance of zero crossovers for each of the models minus .310, the proportion of cases in which zero recombinations occurred among the first five loci in the data. Figures 1b-1d show similar bar graphs for the probabilities of 1,2, and 3 or more crossovers, respectively. The data proportions to which these numbers are compared are .656, .0337, and .000496, respectively.

Figure 2:

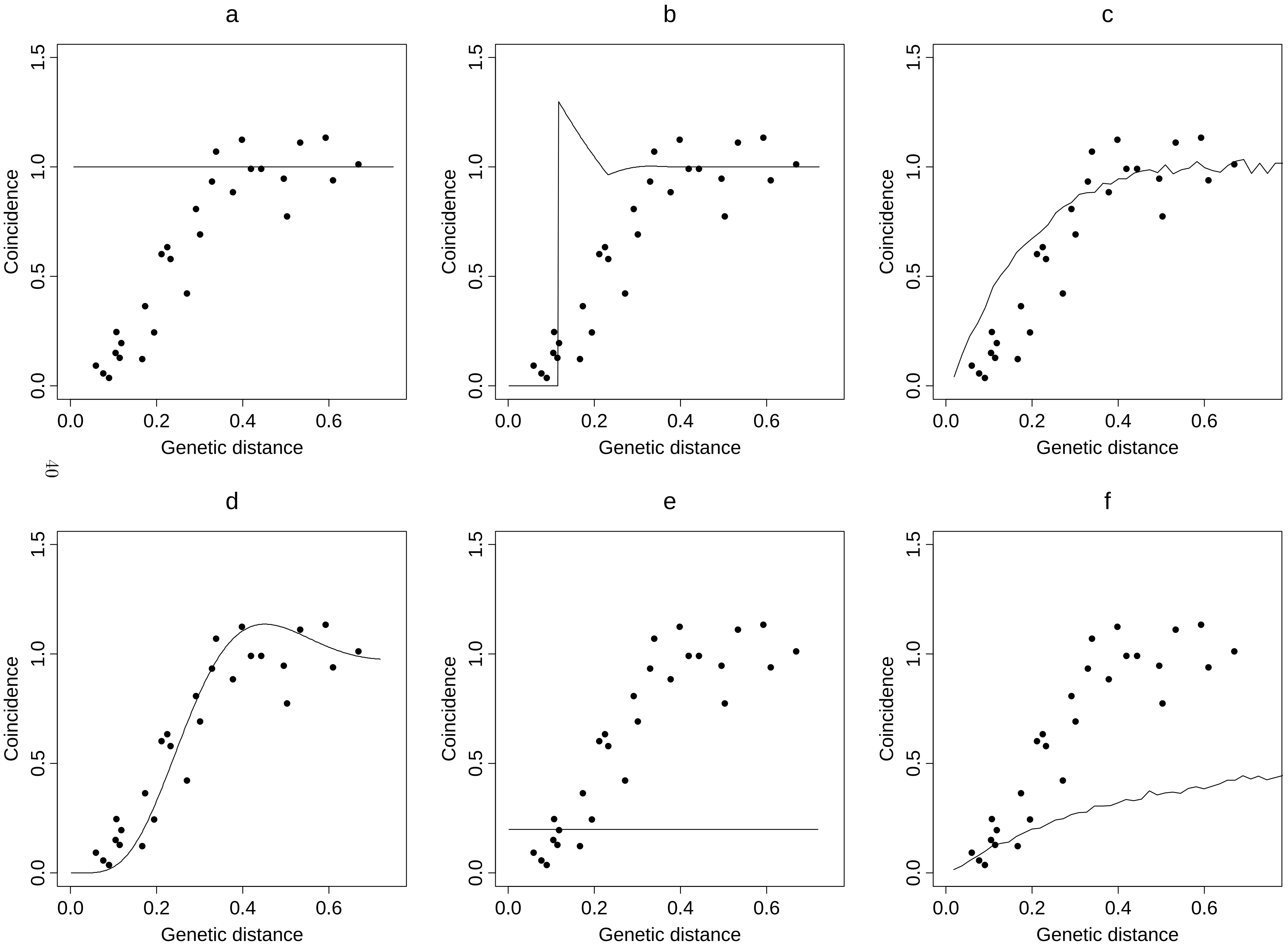
Plots of coincidence vs. genetic distance for the five-locus models. The points represent coincidence calculated from the data for each pair of atomic intervals among the nine loci. The corresponding genetic distance for each pair is calculated between the midpoints of the two intervals. The curves represent five-locus model coincidences. Figure 2a is the Poisson model, 2b is the hard-core model, 2c the King-Mortimer model, 2d the gamma model, 2e the count-location model, and 2f the K-M II model.

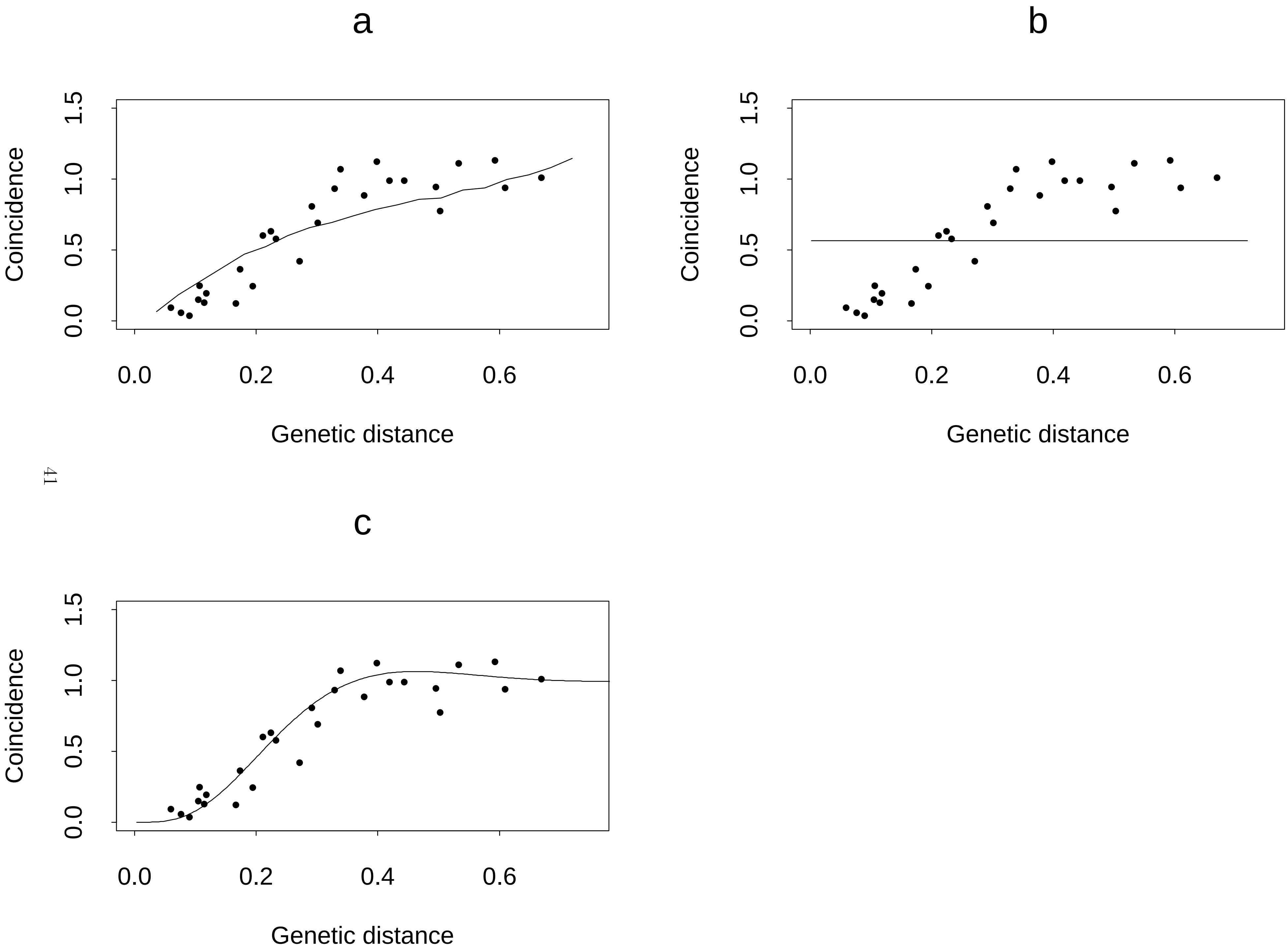
Figure 3:

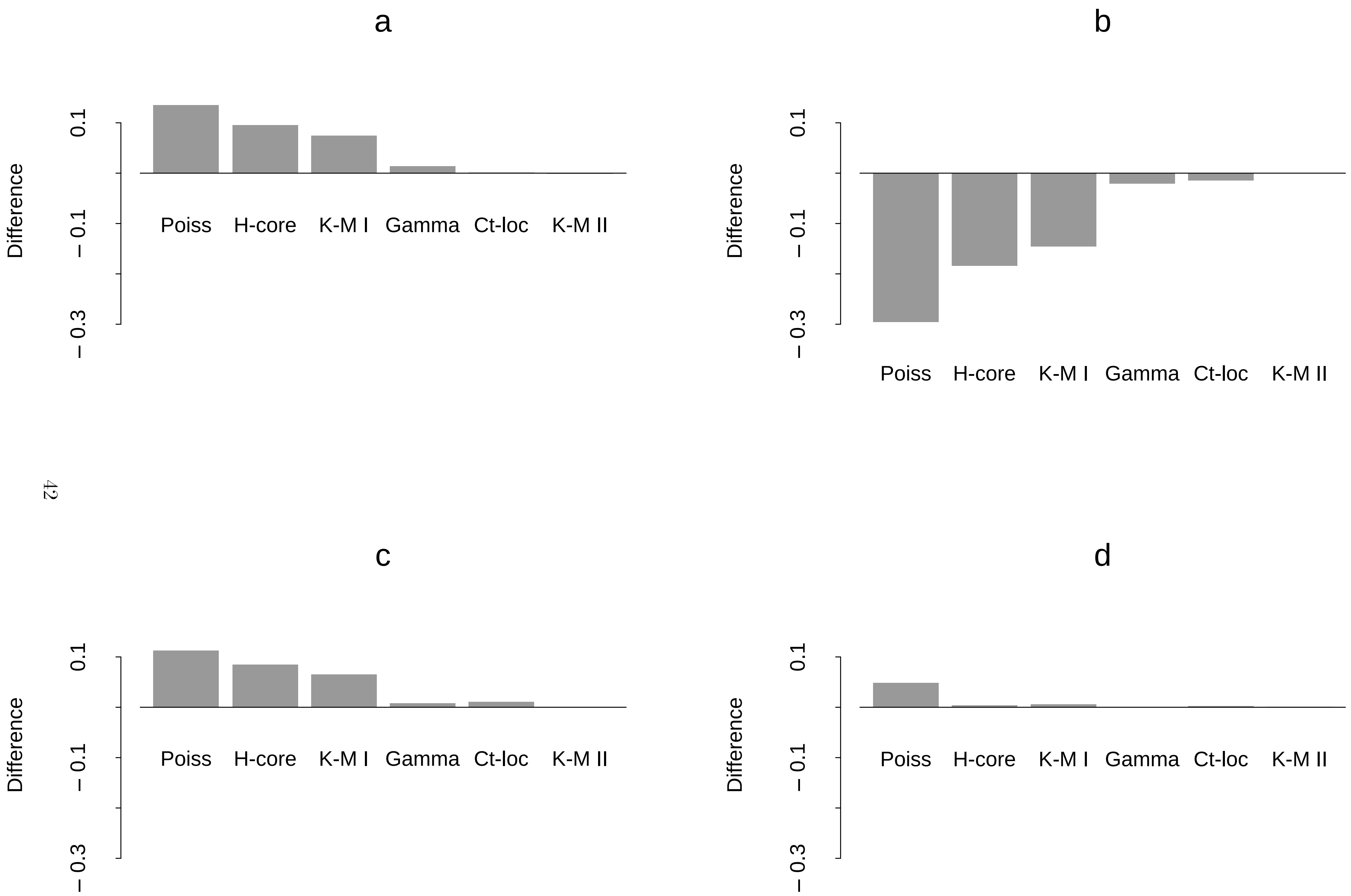
Distribution of the number of crossovers among all nine loci of the MORGAN, BRIDGES and SCHULTZ (1935) dataset, for each of the models under consideration, relative to the data distribution. Figure 3a shows the chance of zero crossovers for each of the models minus .056, the proportion of cases in which zero recombinations occurred in the data. Figures 3b-3e show similar bar graphs for the probabilities of 1,2,3, and 4 or more crossovers, respectively. The data proportions to which these numbers are compared are .485, .429, .030, and 0, respectively.

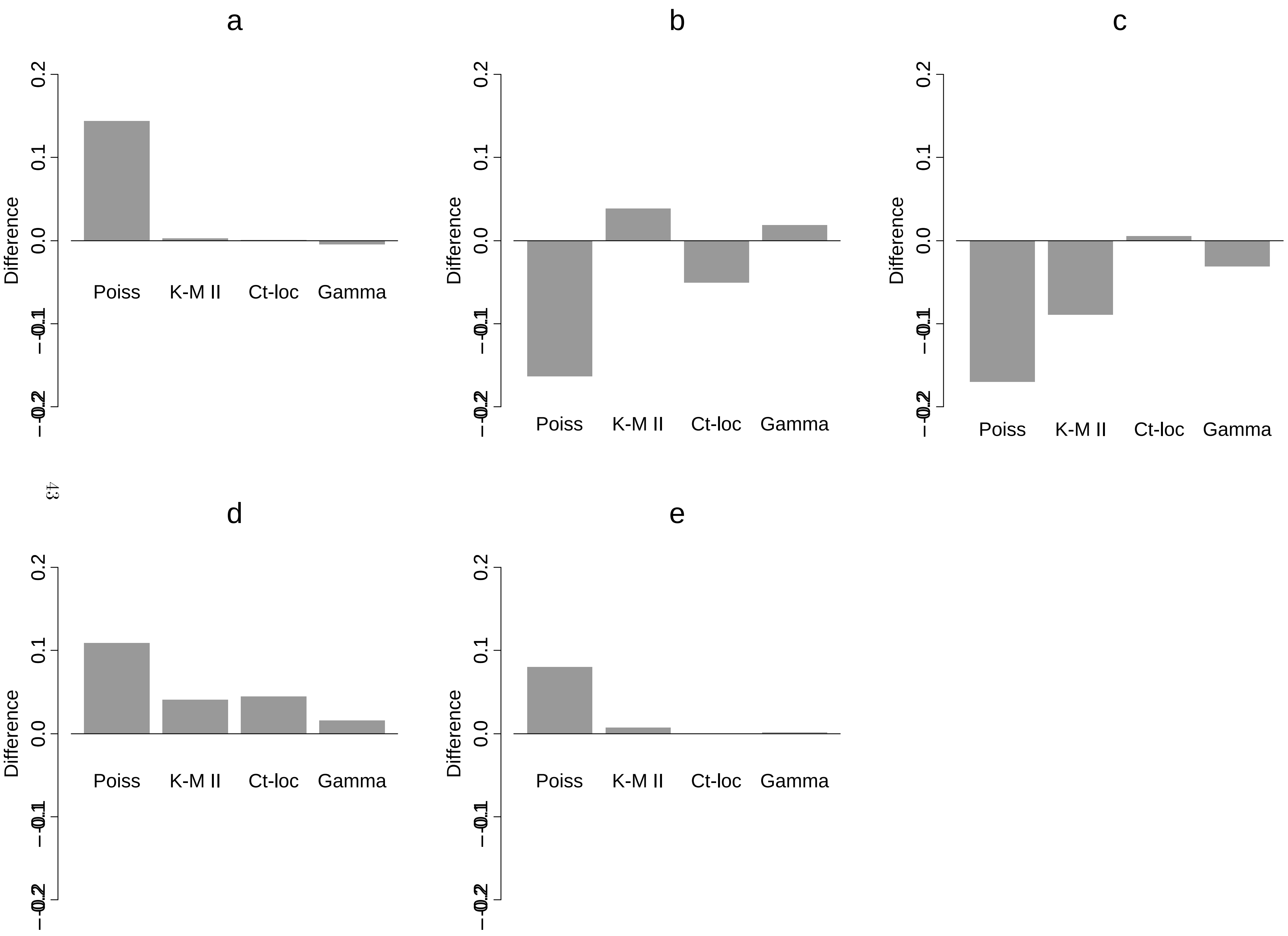
Figure 4:

Plots of coincidence vs. genetic distance for the nine-locus models. The points represent coincidence calculated from the data for each pair of atomic intervals among the nine loci. The corresponding genetic distance for each pair is calculated between the midpoints of the two intervals. The curves represent nine-locus model coincidences. Figure 4a is the K-M II model, 4b is the count-location model, and 4c is the gamma model.


a
b
c
d
e
f
Coincidence
0.0
0.5
1.0
1.5
Genetic distance
0.0
0.2
0.4
0.6


a
b
c
Coincidence
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
Genetic distance


a
Difference
0.1
− 0.1
− 0.3
Poiss
H-core
K-M I
Gamma
Ct-loc
K-M II
b
Difference
0.1
− 0.1
− 0.3
Poiss
H-core
K-M I
Gamma
Ct-loc
K-M II
c
Difference
0.1
− 0.1
− 0.3
Poiss
H-core
K-M I
Gamma
Ct-loc
K-M II
d
Difference
0.1
− 0.1
− 0.3
Poiss
H-core
K-M I
Gamma
Ct-loc
K-M II


a
b
c
d
e
Difference
0.2
0.1
0.0
−0.1
−0.2
Poiss
K-M II
Ct-loc
Gamma