On Genetic Map Functions

Hongyu Zhao* and Terence P. Speed†

**Department of Biostatistics, University of California, Los Angeles, California 90024 and †Department of Statistics, University of California, Berkeley, California 94720*

Manuscript received June 12, 1995
Accepted for publication December 20, 1995

ABSTRACT

Various genetic map functions have been proposed to infer the unobservable genetic distance between two loci from the observable recombination fraction between them. Some map functions were found to fit data better than others. When there are more than three markers, multilocus recombination probabilities cannot be uniquely determined by the defining property of map functions, and different methods have been proposed to permit the use of map functions to analyze multilocus data. If for a given map function, there is a probability model for recombination that can give rise to it, then joint recombination probabilities can be deduced from this model. This provides another way to use map functions in multilocus analysis. In this paper we show that stationary renewal processes give rise to most of the map functions in the literature. Furthermore, we show that the interevent distributions of these renewal processes can all be approximated quite well by gamma distributions.

GENETIC maps consist of different kinds of markers positioned along a chromosome, with their relative locations measured in the map units known as centiMorgans. During meiosis, crossing over occurs after homologous chromosomes pair and duplicate, resulting in a four-strand structure. Each crossover involves two strands of different origins (nonsister pairs). The genetic map distance in Morgans between a pair of markers on the same chromosome is the average number of crossovers occurring between these markers during meiosis on one chromatid. Because crossovers are not observable, a genetic map function $r = M(d)$ is often used to infer genetic distance (d) from the observable recombination fraction (r).

In his 1919 paper, HALDANE made three contributions to the study of genetic map functions, as well as defining them. By assuming no *crossover interference* (STURTEVANT 1915; MULLER 1916), he derived the map function $r = {}^1/_2(1 - e^{-2d})$ with inverse $d = -{}^1/_2 \log(1 - 2r)$; he also proposed the empirical inverse map function $d = 0.7r + 0.3(-{}^1/_2 \log(1 - 2r))$ to account for crossover interference in the data then available, and he introduced a differential equation method that permitted the construction of a variety of map functions.

A commonly used measure of the degree of crossover interference is the coincidence coefficient C involving three markers. Letting $p_{i_1 i_2}$ be the probability of i_1 recombinations in the first interval and i_2 recombinations in the second interval, $i_1, i_2 = 0, 1$, C is defined by

$$C = \frac{p_{11}}{(p_{10} + p_{11})(p_{01} + p_{11})}.$$

Corresponding author: Terence P. Speed, Department of Statistics, University of California, Berkeley, CA 94720.
E-mail: terry@stat.berkeley.edu

The case of $C = 1$ corresponds to no crossover interference, while $C < 1$ and $C > 1$ correspond to *positive* and *negative* crossover interference, respectively. Because the $p_{i_1 i_2}$ can be directly estimated from recombination data in most experimental organisms, C is often used as an empirical index of the degree of crossover interference.

One implicit assumption underlying the use of a genetic map function for one organism is that the functional relationship between genetic distances and recombination fractions does not vary across the genome in this organism. This assumption implies that C should only depend on the map lengths of the intervals between three markers of interest, say d and h. Noting that $p_{10} + p_{11} = M(d)$, $p_{01} + p_{11} = M(h)$, and $p_{11} = \{M(d) + M(h) - M(d + h)\}/2$, we have

$$C(d,h) = \frac{M(d) + M(h) - M(d + h)}{2M(d)M(h)}.$$

Letting $h \to 0$, and assuming $\lim_{h \to 0} M(h)/h = 1$, we obtain the differential equation

$$M'(d) = 1 - 2C(d)M(d). \qquad (1)$$

Different choices of $C(d)$ lead to different map functions, which will be discussed later.

One of the difficulties in using genetic map functions to construct genetic maps is that when there are more than three markers, the multilocus recombination probabilities cannot be uniquely determined from the map function, a point that was made by FISHER (1947). This identifiability problem can be solved by postulating that the probability that there is an odd number of crossovers across a set of (disjoint) intervals only depends on the total length of these intervals

(GEIRINGER 1944; SCHNELL 1961). The underlying assumption is that the distance between two disjoint intervals is irrelevant to the joint recombination probabilities in these two intervals. However experimental results suggest that the degree of interference varies with the distance between two intervals: the smaller the distance, the stronger the interference. Thus, the above approach is not consistent with experimental results. Nevertheless, this assumption can be and has been used to calculate joint recombination probabilities from map functions. LIBERMAN and KARLIN (1984) proved that a necessary condition to guarantee that the recombination probabilities obtained in this way are nonnegative is that $(-1)^k M^{(k)}(0) \leq 0$ for all k. They (inappropriately) termed map functions that always give rise to nonnegative recombination probabilities "multilocus feasible" and showed that many map functions that were found to fit data well did not satisfy these criteria.

If crossovers are viewed as a stochastic point process along the chromatid and a given map function can be realized from a crossover process, multi-locus recombination probabilities compatible with the map function can be obtained by assuming crossovers are generated from this point process. Two classes of point processes have been studied extensively in the literature in the context of modeling crossover interference: renewal processes, reviewed in BAILEY (1961) and MCPEEK and SPEED (1995), and count-location processes, studied by KARLIN and LIBERMAN (1978) and RISCH and LANGE (1979).

In this paper, we show that for most map functions in the literature there exist stationary renewal processes that give rise to them, and so these map functions are compatible with the analysis of multilocus data via this approach. Moreover, the interevent distributions of the stationary renewal processes corresponding to most map functions can be closely approximated by gamma distributions.

A special class of stationary renewal processes, called chi-square models, which have chi-square interevent distributions with even degrees of freedom, was found to give good fit to data from a variety of organisms (ZHAO *et al.* 1995b). This class of models, which evolved from an ordinary renewal process model for crossovers on a single meiotic product proposed by FISHER *et al.* (1947), was mainly studied because of its mathematical tractability (MCPEEK and SPEED 1995). It was reintroduced by FOSS *et al.* (1993) from a biological perspective, motivated by observations from experiments on gene conversion although there are now serious doubts concerning the appropriateness of this motivation (see FOSS and STAHL 1995). The chi-square model was conveniently denoted as $Cx(Co)^m$, which corresponds to a chi-square renewal density with $2(m + 1)$ degrees of freedom. Using the method of maximum likelihood, chi-square models were fitted to different organisms by ZHAO *et al.* (1995b). The estimates of m for Drosophila ($m = 4$) and Neurospora ($m = 2$) were the same as those obtained by FOSS *et al.* (1993), who estimated m from the observed ratio of *Cx* to *Co*. S. LIN and T. P. SPEED (unpublished results) fitted chi-square models to data on six loci from the CEPH consortium map of human chromosome *10*, which was analyzed by WEEKS *et al.* (1994) using other models, and estimated the parameter m to be 3. Their results suggested the presence of crossover interference during human meiosis. MCPEEK and SPEED (1995) compared the fit of the stationary renewal process model with gamma interevent distributions (the gamma model) with that of other models using one large data set of Drosophila and found that the gamma model gives a better fit than all other models. This fact, together with the observation that for most map functions the interevent distributions for the corresponding stationary renewal processes can be closely approximated by gamma distributions, suggest that to some degree the chi-square model, and more generally, the gamma model, is able to capture the important features of the unobservable crossover process.

After raising the question "What is a genetic map function?", SPEED (1995) discussed in detail many issues related to map functions. It should be pointed out that other methods have also been proposed to extend map functions to handle multilocus data (OWEN 1953; MORTON and MACLEAN 1984).

MAP FUNCTIONS AND STATIONARY RENEWAL PROCESSES

When modeling crossovers as a point process, a distinction should be made between the point process on the four strand bundle, *the chiasma process*, and the point process on a single strand, *the crossover process*. Modeling the crossover process as a renewal process has a long history. Following JENNINGS (1923) and MATHER (1936, 1937), FISHER *et al.* (1947) modeled crossing over as a renewal process; that is, crossovers along a single strand were assumed to be formed as a regular sequence starting from the centromere, with the length between two adjacent crossovers always following the same distribution. Although it was known that crossovers take place when four strands (chromatids) are present during meiosis, FISHER *et al.* (1947) only modeled the crossover process on a single strand, without relating it to the chiasma process on the four-strand bundle. On the other hand, most later work modeled the chiasma process as a renewal process and related the crossover process to the chiasma process (CARTER and ROBERTSON 1952; COBBS 1978; STAM 1979). The assumption that crossovers occur starting from the centromere and progressing toward the telomere is now known not to be true (WHITEHOUSE 1982), but it is still a convenience. To connect the chiasma process to the crossover process, assumptions have to be made concerning the way nonsister pairs are involved in crossovers. Two types of

interference can be distinguished: *chiasma interference,* where chiasmata on the four-strand bundle do not occur independently of each other, and *chromatid interference,* where the choices of nonsister pairs involved in different chiasmata are not independent.

The observation of crossover interference on the meiotic products (single strands) can be the result of chiasma interference alone, the result of chromatid interference alone, or the result of both types of interference. It is interesting to note that the operation of two types of interference can lead to no apparent crossover interference. Consider the case when the chiasma process follows a stationary renewal process with interevent distribution being the gamma distribution with shape parameter $^1/_2$, and there is complete positive chromatid interference, *i.e.,* the strands involved in one chiasma are never involved in the closest chiasmata to its left and to its right. It is easy to see that the distance between two crossovers on a single meiotic product from this process follows the gamma distribution with shape parameter 1, *i.e.,* the exponential distribution. Therefore crossovers on a single strand appear to occur independently of each other. This example shows that two types of interference cannot be separated based on single-strand recombination data, where meiotic products from a single meiosis are recovered separately. Therefore tetrad data, where all meiotic products from a single meiosis are recovered together, are often used to detect chromatid interference. Since there is no strong and consistent evidence of chromatid interference (WHITEHOUSE 1982), it is generally assumed that there is no chromatid interference (NCI) in the models proposed in the literature.

The assumption of NCI imposes certain constraints on both recombination and tetrad probabilities (SPEED *et al.* 1992; ZHAO *et al.* 1995a). These constraints further impose constraints on map functions (SPEED 1995):

$$0 \le M(d) \le {}^1/_2,$$
$$M'(d) \ge 0,$$
$$M''(d) \le 0.$$

It is also obvious that if the chiasma point process is simple and stationary in the map distance metric, then $M(0) = 0$ and $M'(0) = 1$ (DALEY and VERE-JONES 1988, Section 3.3). We will consider map functions that are defined on a finite interval $[0, L]$ and those defined $[0, \infty)$ separately. By imposing one more condition, we say a function M defined on $[0, \infty)$ satisfies condition (A) if

$$M(0) = 0, \tag{A1}$$
$$M'(d) \ge 0, \quad \text{for all } d, \tag{A2}$$
$$M'(0) = 1, \tag{A3}$$
$$M''(d) \le 0, \quad \text{for all } d, \tag{A4}$$
$$\lim_{d\to\infty} M(d) = {}^1/_2. \tag{A5}$$

Apart from (A5), these conditions are necessary under the assumptions of NCI and that the chiasma process is a simple stationary point process. Condition (A5) postulates that two markers that are very far apart on the same chromosome can be considered very loosely linked, effectively behaving like markers on different chromosomes and segregating independently. Moreover, we have the following theorem to characterize this class of map functions. The proof is given in the APPENDIX.

Theorem 1: Let M be the map function for a stationary renewal chiasma process satisfying NCI on a chromosome arm of infinite length. Then M satisfies (A). Conversely, suppose that a function M from $[0, \infty)$ into $[0, {}^1/_2]$ satisfies (A). Then there is a stationary renewal chiasma process whose map function is M and whose renewal density is $-M''$.

For a map function M defined on $[0, L]$ where $L < \infty$, we say that M satisfies condition (B) if

$$M(0) = 0, \tag{B1}$$
$$M'(d) \ge 0, \quad \text{for all } d, \tag{B2}$$
$$M'(0) = 1, \tag{B3}$$
$$M''(d) \le 0, \quad \text{for all } d, \tag{B4}$$
$$M'(L) = 0, \tag{B5}$$
$$M(L) = {}^1/_2. \tag{B6}$$

We say that M satisfies condition (B)′ if it satisfies (B1)–(B4) and

$$M'(L) > 0, \tag{B5}'$$
$$M(L) < {}^1/_2. \tag{B6}'$$

For map functions defined on $[0, L]$, we have the following analogue of Theorem 1 with the proof given in the APPENDIX.

Theorem 2: Let M be the map function for a stationary renewal chiasma process satisfying NCI on a chromosome arm of finite length. Then M satisfies (B) or (B)′ for any L. Conversely, suppose that a function M from $[0, L]$ into $[0, {}^1/_2]$ satisfies (B) or (B)′. Then there is a stationary renewal chiasma process whose map function is M and whose renewal density is $-M''$ when $d \le L$.

VARIOUS MAP FUNCTIONS

In this section, we apply our two theorems to some map functions in the literature to see if there are stationary renewal processes that can give rise to them.

HALDANE (1919): $M_{H_2}^{-1}(r) = 0.7r + 0.3(-{}^1/_2 \log(1 - 2r))$. It is easy to see that $M^{-1}(0) = 0$, $\lim_{r\to 1/2} M^{-1}(r) = \infty$, $(M^{-1})' > 0$, $(M^{-1})'(0) = 1$, $\lim_{r\to 1/2} (M^{-1})' = \infty$, and $(M^{-1})'' \ge 0$. So M_{H_2} satisfies condition (A), and there is a stationary renewal process giving rise to M_{H_2}.

LUDWIG (1934): $M_L(d) = {}^1/_2 \sin(2d)$. Clearly this should only be considered a possible map function in

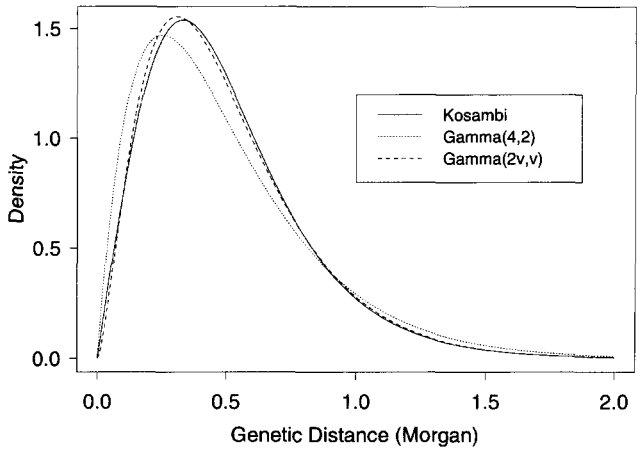

FIGURE 1.—Comparison between the interevent density $(-M_K'')$ of the stationary renewal process corresponding to the Kosambi map function and two gamma densities: gamma(4,2) and gamma(2v,v), where v is $1/(2\log(2) - 1) \approx 2.6$.

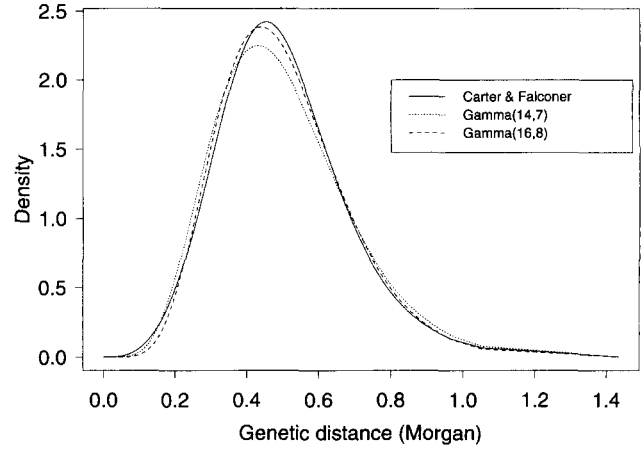

FIGURE 2.—Comparison between the interevent density $(-M_{CF}'')$ of the stationary renewal process corresponding to the Carter and Falconer map function and two gamma densities: gamma(14,7) and gamma(16,8).

the interval $[0, \frac{1}{4}\pi]$, and it is easy to check condition (B) in this case, with $L = \frac{1}{4}\pi$. Thus there is a stationary renewal process having it as a map function, although the chromosome arm is rather short.

KOSAMBI (1944): $M_K(d) = \frac{1}{2}\tanh(2d)$. Since $M_K'(d) = 4(e^{2d} + e^{-2d})^{-2}$, and $M_K''(d) = -16(e^{2d} - e^{-2d})/(e^{2d} + e^{-2d})^3$, it is easy to check that (A) is satisfied. The interevent distribution for the corresponding stationary renewal process is $16(e^{2d} - e^{-2d})/(e^{2d} + e^{-2d})^3$. KOSAMBI (1944) found that this map function gave good fit to Drosophila data. FISHER *et al.* (1947) noticed that the Kosambi map function is very close to the map function from a renewal process with interevent distribution being chi-square with four degrees of freedom. Both $-M_K''$ and the density of gamma(4,2) are plotted in Figure 1. The density of a gamma(b,g) variable is $[b(bx)^{g-1}e^{-bx}]/\Gamma(g)$. The mean and variance of $-M_K''$ are $\frac{1}{2}$ and $\log(2)/2 - \frac{1}{4}$. The density of gamma$(2v, v)$, where $v = 1/(2\log(2) - 1)$, which has the same mean and variance as $-M_K''$, is also plotted in Figure 1.

CARTER and FALCONER (1951): $M_{CF}^{-1}(r) = \frac{1}{4}(\tan^{-1}(2r) + \tanh^{-1}(2r))$. Here $M_{CF}(d)$ is the solution of the differential equation $M'(d) = 1 - 16M^4(d)$. This map function was found to fit mouse data better than other map functions (CARTER and FALCONER 1951). Since $M_{CF}''(d) = -64M_{CF}^3(d)\,(1 - 16M_{CF}^4(d))$, and $M_{CF} \leq \frac{1}{2}$, (A) is satisfied. The corresponding stationary renewal process has interevent distribution $64M_{CF}^3(d)\,(1 - 16M_{CF}^4(d))$. The Carter and Falconer map function $-M_{CF}''(d)$ together with the density of gamma(14,7) and gamma(16,8) are plotted in Figure 2. It was found that $Cx(Co)^6$ and $Cx(Co)^4$ give the best fit for two mouse data sets in BLANK *et al.* (1988) and TODD *et al.* (1991) (ZHAO 1995); the choice of the Carter and Falconer map function would be equivalent to using the $Cx(Co)^6$ model.

STURT (1976): $M_S(d) = \frac{1}{2}(1 - (1 - {}^{d}/_{L})e^{-d(2L-1)/L})$, $0 \leq d \leq L$. This map function arises via a count-location chiasma process that begins with an obligatory crossover event on the arm, followed by a Poisson-distributed number of crossover events having mean $2L - 1$. The total genetic length is thus L. This map function fails (B5) but satisfies (B6), and so no stationary renewal process can give rise to it.

RAO *et al.* (1977): $M_R^{-1}(r) = w_H M_H^{-1}(2r) + w_K M_K^{-1}(2r) + w_{CF}M_{CF}^{-1}(2r) + w_M M_M^{-1}(2r)$, where $w_H = p\,(1 - 2p)(1 - 4p)/3$, $w_K = -4p(1 - p)(1 - 4p)$, $w_{CF} = 32p(1 - p)(1 - 2p)/3$, $w_M = (1 - p)(1 - 2p)(1 - 4p)$, and M_H^{-1}, M_K^{-1}, M_{CF}^{-1}, and M_M^{-1} are the inverse of the Haldane, Kosambi, Carter-Falconer, and Morgan map functions. It is easy to check that for any p, w_H, w_K, w_{CF}, and w_M cannot all be positive. Indeed M_R does not satisfy our necessary conditions (WEEKS 1994). Following RAO *et al.*'s idea, we might try to define a map function from a set of $n > 2$ map functions by letting $M^{-1}(r) = \sum_{i=1}^{n} w_i(p)\, M_i^{-1}(r)$, where $w_i(p)$ is a polynomial in p of order $n - 1$, and $M(d)$ reduces to $M_i(d)$ when $p = p_i$ for given $0 < p_1 < p_2 < \cdots p_n < 1$. But it can be shown that for no p can the $w_i(p)$, $i = 1, 2, \ldots, n$, so defined all be positive. Therefore this approach to obtaining empirical map functions from existing map functions would seem questionable.

FELSENSTEIN (1979): $M_F(d) = \frac{1}{2}\,(1 - e^{2(K-2)d})/(1 - (K-1)e^{2(K-2)d})$. When $K > 2$, $\lim_{d\to\infty} M_F(d) = \frac{1}{2}\,(K - 1) \neq \frac{1}{2}$, and so no stationary renewal process exists with this mapping function. Because $M_F'(d) = (K - 2)^2 e^{2(K-2)d}/(1 - (K-1)e^{2(K-2)d})^2$, $M_F''(d) = 2\,(K - 2)^3 e^{2(K-2)d}\,(1 + (K-1)e^{2(K-2)d})/(1 - (K-1)e^{2(K-2)d})^3$, it is easy to verify that $M_F(d)$ satisfies (A) when $0 \leq K < 2$. In Felsenstein's map functions, K is a measure of interference, with $K > 1$ corresponding to negative interference, and $K < 1$ to positive interference. The mean of $-M_F''(d)$ is $\frac{1}{2}$, and the variance is $\log(2 - K)/(2(1 - K)) - \frac{1}{4}$ ($\frac{1}{2}$ when $K = 1$). Both $-M_F''(d)$ and the gamma distribution with the same mean and variance are plotted in Figure 3 for different Ks. Note the

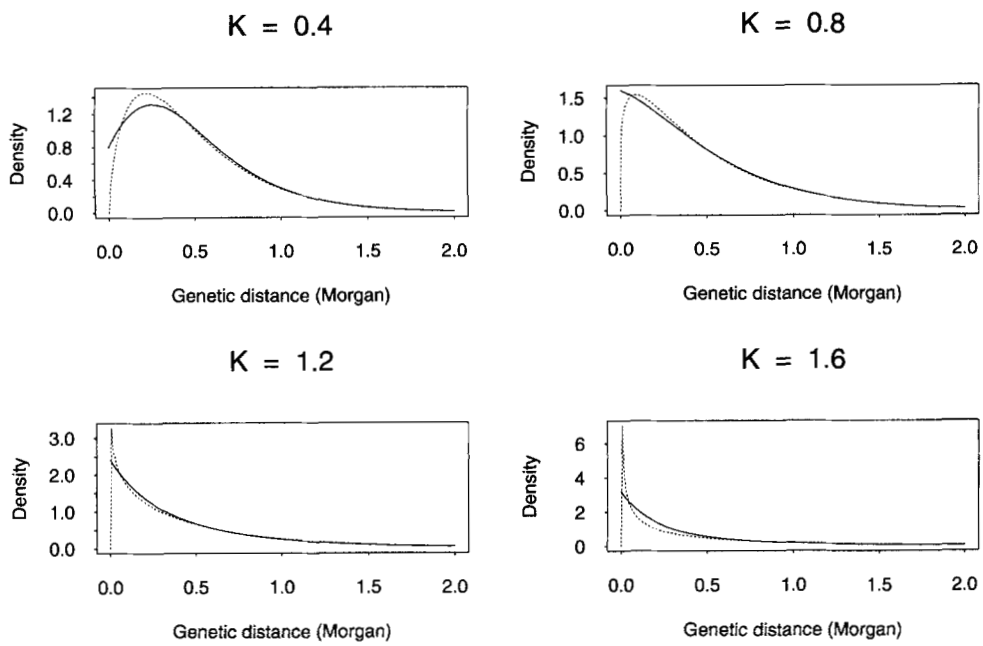

FIGURE 3.—Comparison between the interevent density ($-M_K'$, solid curve) of the stationary renewal process corresponding to the Felsenstein map function and the gamma density (dotted curve) having the same mean and variance for $K = 0.4$, 0.8, 1.2, and 1.6.

close agreement between Felsenstein's family and the gamma family.

KARLIN and LIBERMAN (1978, 1979): $M_{CL}(d) = {}^1/_2\,[1 - c(1 - d/L)]$. This class of map functions arise from the count-location process, where $c(s) = \sum_{k\geq 0} c_k s^k$ is a probability generating function of a count variable **c** with distribution (c_k), and $c'(1) = 2L$. M_{CL} is only well defined for finite L, It is easy to check that (B1)–(B4) are satisfied. Because $M_{CL}(L) = {}^1/_2\,(1 - c_0)$ and $M'_{CL}(L) = c_1/2L$, from Theorem 2, there is a corresponding stationary renewal process for M_{CL} only if (1) $c_0 = 0$ and $c_1 = 0$, or (2) $c_0 > 0$ and $c_1 > 0$.

DISCUSSION

In constructing genetic maps, map functions have been used to infer the unobservable genetic distance between two markers from the observable recombination fraction between these markers. Different genetic map functions embody different degrees of crossover interference among the crossovers. It has been observed that different organisms have different degrees of crossover interference, it so is not surprising that different map functions have been found suitable for different organisms. The major disadvantage of using map functions is that joint recombination probabilities cannot be uniquely determined in terms of them when there are more than three markers. Various approaches have been proposed to extend map functions to handle multilocus data.

One widely adopted approach, which was suggested by GEIRINGER (1944) and SCHNELL (1961) and thoroughly studied by LIBERMAN and KARLIN (1984), embodies the assumption that for a pair of noncontiguous intervals, the probabilities for joint recombination patterns across these intervals do not depend on the distance between these two intervals, something that is not consistent with observations. Those map functions that can be extended to multilocus data through this approach have been (inappropriately) called "multilocus feasible" (LIBERMAN and KARLIN 1984). This criterion excludes many functions that were found to fit well to recombination data, such as the Kosambi map function.

In this paper, another approach is proposed to extend map functions for the analysis of multilocus data. If for any given map function, we can find a point process model that gives rise to this map function, then multilocus joint recombination probabilities can be obtained in a way that is completely compatible with the map function. Stationary renewal processes can give rise to many map functions. From this perspective, most map functions that are not multilocus feasible according to KARLIN and LIBERMAN can in fact be extended to permit the analysis of multilocus data.

Another measure of interference, called S_4 by FOSS *et al.* (1993), is formally defined as

$$S_4(d) = \lim_{h\to 0}\lim_{k\to 0} \frac{p_{11}}{(p_{10} + p_{11})(p_{01} + p_{11})},$$

where the $p_{i_1 i_2}$ are as in the definition of C, with one interval having map length h, the other interval map length k, and the two intervals being separated by a map distance d. It seems that S_4 captures more important aspects of crossover interference than does C (MCPEEK and SPEED 1995). Though S_4 cannot be deter-

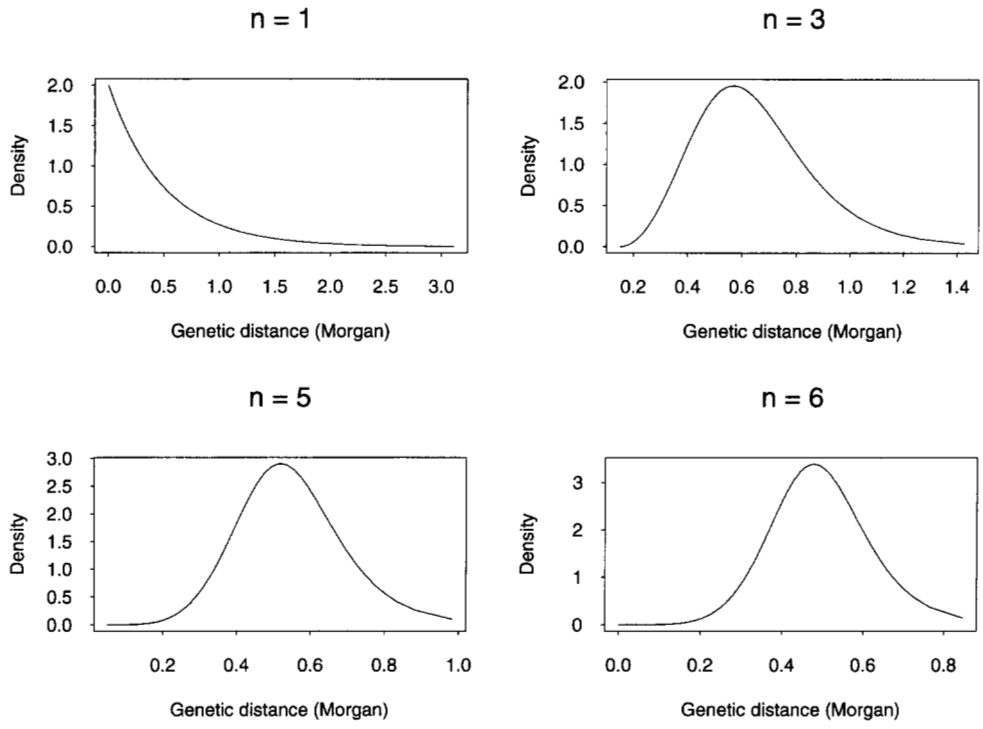

FIGURE 4.—The interevent density of the stationary renewal process corresponding to the map function M that satisfies $M' = 1 - (2M)^n$ for $n = 1, 3, 5, 6$.

mined from a map function, it can be calculated given a particular way of generalizing map functions to multilocus data and compared to empirical estimates. For the count-location processes, S_4 is constant for all d, whereas S_4 has various forms for the stationary renewal processes. The values of S_4 as a function of d were estimated from a large Drosophila data set and were very close to the S_4 values under the chi-square model (FOSS *et al.* 1993, MCPEEK and SPEED 1995).

Map functions cannot and should not be expected to reflect chiasma or crossover interference in anything but the most superficial way. Indeed, a map function could arise from both the stationary renewal process and the count-location process, although the interference differs greatly between these two classes of models. For example, the map function $M(d) = d/(1 + 2d)$ could arise from a stationary renewal process with interevent distribution $4(1 + 2d)^{-3}$, and it could also arise from a count-location model with $c_k = {}^1/_2{}^k$, where $k \geq 0$ (LIBERMAN and KARLIN 1984). Under the count-location model, the length of the chromosome defined by $\mathbf{c} = (c_k)$ is ${}^1/_2$, so $M(d)$ is only defined on $[0, {}^1/_2]$, while d can range from 0 to ∞ for the stationary renewal process model.

Note that for the stationary renewal processes corresponding to most map functions discussed, the interevent distributions can often be well approximated by gamma distributions. Recall that the Haldane, Kosambi, and Carter and Falconer map functions are all solutions to the differential Equation 1 with $C(d) = (2M)^{n-1}$. It can be shown that map functions M obtained via this approach always satisfy (A). Figure 4 displays plots of $-M''(d)$, the interevent distribution of the corresponding stationary renewal process for $n = 1$ (Haldane map function), $n = 3$, 5, and 6. The cases $n = 2$ and 4, which correspond to the Kosambi map function and the Carter and Falconer map function, were plotted in Figure 1 and Figure 2. When n is large, the corresponding density becomes close to a normal distribution. This suggests that if we take the approach proposed in this paper to generalize map functions to the multilocus situation, the stationary renewal processes corresponding to most genetic map functions can be well approximated by the gamma model, or the chi-square model. This provides, to some extent, an explanation of the fact that chi-square models (and hence gamma models) are flexible enough to give good fit to data from various organisms exhibiting different degrees of interference.

It has been assumed throughout this paper that there is no chromatid interference. Under this assumption, the relation between the chiasma process and the crossover process and the relation between the chiasma process and the map function are simple [see SPEED (1995) for a comprehensive review of this and related matters]. It was shown earlier that there are situations where the presence of both chiasma interference and chromatid interference could lead to no apparent crossover interference, from which we concluded that it is impossible to separate two types of interference from single-strand recombination data. Because there is no strong and consistent evidence of chromatid interference from the study of tetrad data, the assumption of no chromatid

interference is generally considered valid. The use of map functions and stationary renewal processes requires that the degree of interference be the same across the chromosome, which is an obvious simplification. But a large amount of data will probably be necessary to permit the detection of nonstationarity of the underlying process.

In summary, we have shown that for most genetic map functions, there is a corresponding stationary renewal process, and that these map functions can be extended to permit the analysis of multilocus data by calculating joint recombination probabilities from their corresponding renewal processes. This provides another way of generalizing a given map function to the multilocus situation, although it seems unlikely that there are efficient methods to estimate genetic distances and other parameters from multilocus recombination data under completely general renewal processes. The calculation of multilocus recombination probabilities for stationary renewal chiasma processes is discussed in the APPENDIX. However, comparisons between the interevent distribution of general stationary renewal processes and that of chi-square distributions suggest that this class of renewal processes provides satisfactory approximations to the renewal processes corresponding to most genetic map functions in the literature. With the limited amounts of data currently available, it will probably be hard to distinguish these models, although we can look forward to many refinements in the future.

This work was supported by National Institutes of Health grant HG-01093-01. The authors thank two referees for their helpful comments.

LITERATURE CITED

BAILEY, N. T. J., 1961 *Introduction to the Mathematical Theory of Genetic Linkage.* Oxford University Press, London.

BLANK, R. D., G. R. CAMPBELL, A. CALABRO and P. D. EUSTACHIO, 1988 A linkage map of mouse chromosome 12: localization of *Igh* and effects of sex and interference on recombination. Genetics **120:** 1073–1083.

CARTER, T. C., and D. S. FALCONER, 1951 Stocks for detecting linkage in the mouse and the theory of their design. J. Genet. **50:** 307–323.

CARTER, T. C., and A. ROBERTSON, A., 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.

DALEY, D. J., and D. VERE-JONES, 1988 *An Introduction to the Theory of Point Processes.* Springer-Verlag, New York.

FELLER, W., 1971 *An Introduction to Probability Theory and Its Applications.* Vol. 2, Ed. 2. John Wiley, New York.

FELSENSTEIN, J., 1979 A mathematically tractable family of genetic mapping functions with differents amount of interference. Genetics **91:** 769–775.

FISHER, R. A., 1947 The theory of linkage in polysomic inheritance. Phil. Trans. Roy. Soc. B. **233:** 55–87.

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., and F. W. STAHL, 1995 A test of a counting model for chiasma interference. Genetics **139:** 1201–1209.

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

GEIRINGER, H., 1944 On the probability theory of linkage in Mendelian heredity. Ann. Math. Statist. **15:** 25–57.

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.

JENNINGS, H. S., 1923 The numerical relations in the crossing over of the genes with a critical examination of the theory that the genes are arranged in a linear series. Genetics **8:** 393–457.

KARLIN, S., and U. LIBERMAN, 1978 Classification and comparisons of multilocus recombination distributions. Proc. Natl. Acad. Sci. USA **75:** 6332–6336.

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

KOSAMBI, D. D., 1944 The estimation of the map distance from recombination values. Ann. Eugen. **12:** 172–175.

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

LUDWIG, W., 1934 Uber numerische beziehungen der crossoverwerte untereinander. Z. indukt. Abatamm. Vereb. **67:** 58–95.

MATHER, K., 1935 Reductional and equational separation of the chromosomes in bivalents and multivalents. J. Genet. **30:** 53–78.

MATHER, K., 1936 The determination of position in crossing over. J. Genet. **33:** 207–235.

MATHER, K., 1937 The determination of position in crossing over. II. The chromosome length-chiasma frequency relation. Cytologia, Jub. Vol., 514–526.

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

MORGAN, T. H., and C. B. BRIDGES, 1916 Sex-linked inheritance in Drosophila. Carniegie Institute of Washington.

MORTON, N. E., and C. J. MACLEAN, 1984 Multilocus recombination frequencies. Genet. Res. **44:** 99–108.

MULLER, H. J., 1916 The mechanism of crossing over. Am. Nat. **50:** 193–221, 284–305, 350–366, 421–434.

OWEN, A. R. G., 1953 The analysis of multiple linkage data. Heredity **7:** 247–264.

RAO, D. C., N. E. MORTON, J. LINDSTEN, M. HULTEN, and S. YEE, 1977 A mapping function for man. Hum. Hered. **27:** 99–104.

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

SCHNELL, F. W., 1961 Some general formulations of linkage effects in inbreeding. Genetics **46:** 947–957.

SPEED, T. P., 1995 What is a genetic map function? in *Genetic Mapping and DNA Sequencing,* edited by T. P. SPEED and M. S. WATERMAN. Springer-Verlag, New York.

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

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

STURT, E., 1975 A mapping function for human chromosomes. Ann. Hum. Genet. **40:** 147–163.

STURTEVANT, A. H., 1915 The behavior of chromosomes as studied through linkage. Z. Indukt. Abstammungs. Vererbugsl. **13:** 234–287.

TODD, J. A., T. J. AITMAN, R. J. CORNALL, S. GHOSH, J. R. HALL *et al.,* 1991 Genetic analysis of autoimmune type 1 diabetes mellitus in mice. Nature **351:** 542–547.

WEEKS, D. E., 1994 Invalidity of the Rao map function for three loci. Hum. Hered. **44:** 178–180.

WEEKS, D. E., G. M. LATHROP and J. OTT, 1993 Multipoint mapping under genetic interference. Hum. Hered. **43:** 86–97.

WHITEHOUSE, H. L. K., 1982 *Genetic Recombination: Understanding the Mechanisms.* John Wiley, New York.

ZHAO, H., 1995 *Statistical Analysis of Genetical Interference.* PhD thesis, University of California at Berkeley, Berkeley.

ZHAO, H., M. S., MCPEEK and T. P. SPEED, 1995a Statistical analysis of chromatid interference. Genetics **139:** 1057–1065.

ZHAO, H., T. P. SPEED and M. S. MCPEEK, 1995b Statistical analysis of crossover interference using the chi-square model. Genetics **139:** 1045–1056.

Communicating editor: B. S. WEIR

APPENDIX

Proof of Theorem 1: Suppose the crossovers are from a stationary renewal chiasma process with interevent density f. Without loss of generality, we may assume the mean interevent distance is $\mu = {}^1/_2$, so the metric is simply genetic distance. For any point $\mathcal{M}_1$, say, on the chromosome, the chance that the first crossover after $\mathcal{M}_1$ occurs in the small interval $(y, y + dy)$ is

$$\int_y^\infty \frac{f(t)}{\mu}\, dtdy = 2\int_y^\infty f(t)\, dtdy.$$

The probability of no crossovers occurring before $\mathcal{M}_2$, which is map distance d from $\mathcal{M}_1$, is

$$p_0 = \int_d^\infty \left\{2\int_y^\infty f(t)\, dt\right\} dy.$$

Using Mather's formula (MATHER 1935), which asserts that the recombination fraction r between any two markers is

$$r = \frac{1}{2}(1 - p_0),$$

where p_0 is the probability of zero crossovers occurring between these markers, we have

$$M(d) = r = \frac{1}{2}\left\{1 - 2\int_d^\infty \int_y^\infty f(t)\, dtdy\right\}.$$

It is easy to verify that M so defined satisfies (A).

Conversely, if M satisfies (A), then

$$\int_0^\infty (-M''(t))\, dt = M'(0) = 1.$$

Thus $-M''$ is a probability density function on $[0, \infty)$. If the interevent distribution in a stationary renewal process is $-M''$, then their mean is

$$\int_0^\infty (-tM''(t))\, dt = \int_0^\infty \int_t^\infty (-M''(y))\, dydt$$
$$= \int_0^\infty M'(t)\, dt = {}^1/_2.$$

Thus, the map function generated from the stationary renewal chiasma process with interevent distribution $-M''$ is

$$\frac{1}{2}(1 - 2\int_d^\infty \int_y^\infty (-M''(t)\, dt)\, dy) = M(d).$$

Proof of Theorem 2: Note that

$$M(L) = \frac{1}{2}\left\{1 - 2\int_L^\infty \int_y^\infty f(t)\, dtdy\right\},$$

and

$$M'(L) = \int_L^\infty f(t)\, dt.$$

It follows that either (B5) and (B6) are true, or (B5)' and (B6)' are true. The first part of this theorem can then be proved as in Theorem 1.

If $M(L) < {}^1/_2$ and $M'(L) > 0$, we may define an extended map function $M_E(d)$ on $[0, \infty)$ as follows:

$$M_E(d) = \begin{cases} M(d) & \text{if } d < L \\ M(L) + \alpha(1 - e^{-\beta(d-L)}) & \text{if } d \geq L, \end{cases}$$

where $\alpha = {}^1/_2 - M(L)$ and $\alpha\beta = M'(L)$. It can be verified that $M_E(d)$ so defined satisfies (A) in Theorem 1. So there is a stationary renewal process whose corresponding map function is $M_E(d)$, which coincides with $M(d)$ on $[0, L]$. If $M(L) = {}^1/_2$ and $M'(L) = 0$, it can be easily shown that the stationary renewal process with renewal density $-M''$ gives rise to M.

Calculating multilocus recombination probabilities for stationary renewal chiasma processes: Suppose that $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_n$ are $n + 1$ consecutive loci along a chromosome, defining n genomic intervals $I_1 = [\mathcal{M}_0, \mathcal{M}_1)$, $\ldots$, $I_n = [\mathcal{M}_{n-1}, \mathcal{M}_n)$ of map lengths $d_1, d_2, \ldots, d_n$, respectively. Extending the notation introduced earlier, we denote by $p_{i_1 \ldots i_n}$ the joint recombination probability of having $i_j = 0$ or 1 recombination across interval I_j, $j = 1, \ldots, n$. The question we address here is the calculation of all such probabilities $\mathbf{p} = (p_{i_1 \ldots i_n})$ when the underlying chiasmata form a renewal process stationary in the genetic distance metric and NCI is assumed. To do so we make use of the so-called *linkage values*, denoted by $\mathbf{z} = (z_{i_1 \ldots i_n})$, where $z_{i_1 \ldots i_n}$ is the probability of finding no chiasmata in the union $\cup\{I_j; i_j = 1\}$ of those intervals for which $i_j = 1$, see SPEED *et al.* (1992) for fuller details. We also make use of some well known facts from renewal theory and refer to FELLER (1971) for derivations. Suppose that we have a stationary renewal process with interevent density f and mean interval length μ. The distance of an arbitrary but fixed point on the chromosome to the next chiasma to its left (respectively right) is called the backward (respectively forward) recurrence length (traditionally called "time"), and if these are denoted by β and ϕ, then

$$P(\beta > u, \phi > v) = \int_{u+v}^\infty \frac{1 - F(y)}{\mu}\, dy, \qquad (2)$$

where F is the cumulative distribution function (*c.d.f.*) corresponding to f.

Now let us consider the calculation of the linkage values $\mathbf{z} = (z_{i_1 \ldots i_n})$. For $n = 2$ this is quite straightforward. Suppose that we want to calculate z_{10}, the probability of no crossovers in I_1. We regard $\mathcal{M}_1$ as the arbitrary but fixed point in the preceding discussion, and put $u = d_1$ and $v = 0$ in (2), obtaining the formula $z_{10} = 1 - F^*(d_1)$, where F^* is the *c.d.f.* corresponding to $f^* = \mu^{-1}(1 - F)$. Similarly, $z_{01} = 1 - F^*(d_2)$, $z_{11} = 1 - F^*(d_1 + d_2)$, and $z_{00} = 1$. All of these expressions are easily computed as long as F^* is tractable.

We now turn to $n = 3$ intervals and the calculation of

$\mathbf{z} = (z_{i_1 i_3})$. A quick run through all eight possibilities reveals that all but z_{101} can be obtained in the manner just illustrated with $n = 2$. For example, $z_{100} = 1 - F^*(d_1)$, $z_{011} = 1 - F^*(d_2 + d_3)$, etc. It turns out that the calculation of z_{101}, in general requires summing a series of multiple integrals, and that the only known cases in which these integrals simplify into something tractable are variants on thinned Poisson processes. Let us see why.

First we recall that z_{101} is the probability of no chiasmata in either I_1 or I_3; there may be zero, one or more in I_2, where the count is not constrained. Thus an initial reduction of z_{101} is as follows:

$$z_{101} = z_{111} + \sum_{k=1}^{\infty} \zeta_k,$$

where z_{111} is the probability of no chiasmata in $I_1 \cup I_2 \cup I_3$ $(= 1 - F^*(d_1 + d_2 + d_3))$, and ζ_k is the probability of k chiasmata in I_2 and none in I_1 or I_3. We now give an expression for ζ_k that, in general, does not simplify, and we remark that we know of no substantially simpler alternative expressions in the literature on renewal processes.

If there are to be $k \geq 1$ chiasmata in I_2, we may denote the forward recurrence interval length from $\mathcal{M}_1$ to the first event by y_1, and the k subsequent interevent distances by $y_2, y_3, \ldots, y_k, y_{k+1}$. Further we may denote by y_0 the backward recurrence distance to the first event to the left of $\mathcal{M}_1$. With this notation we can readily check that the probability ζ_k is the $(k + 2)$-fold integral of the joint density of $(y_0, y_1, \ldots, y_k, y_{k+1})$ over the range $\{y_0 > d_1\} \cap \{y_1 + \cdots + y_k < d_2\} \cap \{y_1 + \cdots + y_k + y_{k+1} > d_2 + d_3\}$. The joint density of $y_0, \ldots, y_{k+1}$ is the product

$$\mu^{-1} f^*(y_0 + y_1) \times \prod_{i=2}^{k+1} f(y_i),$$

and so our assertion is demonstrated: z_{101} is an infinite sum of multiple integrals and will have a tractable expression only when these sums and integrals simplify. For each $n \geq 3$ there is one or more $z_{i_1 \ldots i_n}$ requiring such expressions, and so far it is only the class of chi-square renewal processes (ZHAO *et al.* 1995b) and a slightly more general class termed Poisson-skip processes (H. ZHAO, K. LANGE and T. P. SPEED, personal communication) that are known to yield simplifications.