Genetic Map Functions

A genetic map function M gives a relation $r = M(d)$ connecting recombination fractions r and genetic map distances d between pairs of loci along a chromosome arm. Recombination fractions and map distances are summary statistics concerning potentially observable characteristics of the single chromosomes (also known as chromatids) that are the products of meiosis, and that go into gametes. The recombination fraction between two loci is the proportion of such chromosomes that are recombinant, that is, that have genetic material of differing parental origins, at the two loci (*see* **Linkage Analysis, Model Based**). The genetic map distance between two loci is the average number of exchange points that occur along such a chromosome between the loci, where an exchange point, also known as a crossover point, is a point where the parental origin of the genetic material changes. In these definitions, proportions and averages are calculated in the hypothetical infinite population of single chromosomes resulting from meiosis in a given organism, occurring under standard conditions. Variations between organisms within the same species, or of the conditions of meiosis, may lead to small, but observable, differences in these quantities. It should be noted that some authors (e.g. [1] and [9]) use the term map function for the function M^{-1} in the inverse relation $d = M^{-1}(r)$ expressing d in terms of r. We follow Karlin [7] and others in calling M a map function, mainly because the theoretical development is slightly simpler for M than for M^{-1}.

Map functions have been widely used in genetics because of two facts. The first is that genetic map distances are additive by definition, whereas recombination fractions are not. Thus, map distances are preferred for mapping chromosomes. The second is that recombination fractions are much easier to estimate from data, although with human data indirect techniques may need to be used, see [9]. This is because recombination refers only to features of chromosomes at the endpoints of intervals. By contrast, to estimate a map distance information concerning exchanges in the entire interval between two loci is required and, until recently, such information was rarely, if ever, available. Modern molecular genetic methods now exist permitting the identification of points of exchange along chromosomes, and in the near future it may become much easier to estimate map distances directly (see [8]).

The traditional use of map functions has been to take an estimated recombination fraction $\hat{r}$ between two loci and a map function M deemed appropriate for the organism in question, and estimate the map distance between the loci by the quantity $\hat{d} = M^{-1}(\hat{r})$. Perhaps the simplest case is the map function $r = d$, with inverse $d = r$. This is quite satisfactory for small r and d, say, in the interval (0, 0.05), but the relative error increases as the magnitudes of d and r increase. If two loci can be linked by a chain of intermediate loci, each having a recombination fraction of no more than 0.05 (say) with its successor, then a quite satisfactory estimate of the map distance between the initial and final locus can be obtained by adding the successive interlocus recombination fractions. The notion of map function is helpful in situations where such intermediate loci are not available.

The recombination fraction and map length of an interval will differ when there is a nonzero chance of multiple exchange points occurring in the interval. The chance of this occurring increases as the size of the interval increases. If we denote the distribution of exchange points in a particular interval by $(p_0, p_1, p_2, p_3, \ldots)$, so that p_k is the expected proportion of single chromosomes that have k exchange points in the interval, then the recombination fraction is

$$r = p_1 + p_3 + \cdots \tag{1}$$

(i.e. the probability of an odd number of exchange points), while the map length is

$$d = p_1 + 2p_2 + 3p_3 + \cdots. \tag{2}$$

For example, if $p_k = \mathrm{e}^{-d}d^k/k!$, then the map length is easily seen to be d, while the recombination fraction is

$$r = \mathrm{e}^{-d} + \mathrm{e}^{-d}\frac{d^3}{3!} + \cdots = \tfrac{1}{2}(1 - \mathrm{e}^{-2d}). \tag{3}$$

This relation is known as Haldane's map function, and it is widely used today, nearly 80 years after Haldane [6] first described it. Although simple and easy to use, especially for multilocus calculations,

Reproduced from the *Encyclopedia of Biostatistics, 2nd Edition*. © John Wiley & Sons, Ltd. ISBN: 0-470-84829-4.

the **Poisson Process** underlying this map function has only rarely been found to fit recombination data. As a result, a sizeable body of work in the late 1940s and 1950s from R.A. Fisher and colleagues and students, excellently summarized in [1], supposed that the points of exchange along a chromosome follow a **renewal process** with independent interpoint distances distributed as $\frac{1}{4}\chi_4^2$ or $\frac{1}{6}\chi_6^2$, rather than the $\frac{1}{2}\chi_2^2$ that gives rise to Haldane's map function. Their model seemed to fit existing human, mouse, and other data quite satisfactorily, but possesses no map function.

The very notion of a map function embodies certain implicit biological assumptions about the process of recombination. For example, all map functions in the literature are bounded above by 1/2, thereby constraining recombination fractions to be $\leq$1/2. This is widely believed to hold, but there have been instances where it was felt to be untrue, see [4]. Less obviously, the use of a map function presumes that distinct chromosomal intervals having the same map length necessarily have the same recombination fraction, and conversely. This form of stationarity or homogeneity is not observed in the one case in which there is enough data to test it [3]. A number of writers have discussed probability models for recombination that do not constrain recombination fractions to be $\leq$1/2, and do not satisfy the stationarity properties leading to a map function, see [1] and [5]. Map functions are best viewed as an aspect of certain probability models for recombination. As such, they reflect modeling assumptions, and cannot be expected to be consistent with all the relevant biological knowledge. What matters is whether they are effective for the purposes to which they are put.

Map functions are also useful in contexts where all the products of meiosis remain together, as is the case with ordered or unordered tetrads or octads. In such situations, the model needs to be modified slightly, for although the concept of map distance remains appropriate, the classification of chromosomes as recombinant or not between loci is replaced by a classification of tetrads or octads depending on the parental origins of genetic material at the loci (see, for example, [2]). We will not give any details, here, but simply observe that this development leads us to consider probability models for recombination that refer to the four-strand bundle of chromatids, rather than to the single chromosome products of meiosis. In this approach, chiasmata (the chromosomal structures at points of exchange) are postulated to occur along the four-strand bundle according to some point process, and a mechanism for determining the strands involved in the chiasmata is also specified. The distribution of change-points along the resulting chromosomes is then a consequence of the interplay between the chiasma location process and the strand choice mechanism, and, in specifying the recombination process in this manner, we are also able to calculate the probabilities of interest concerning tetrad and octad types. The simplest assumption concerning strand choice is that the strands involved in any given chiasma are chosen at random from the four possible, independently of those chosen for other chiasmata. This is known as the assumption of no chromatid interference, interference being a term used in genetics to denote some form of dependence. In what follows we make this assumption, although (see [11]) map functions can be defined without it.

Under the assumption of no chromatid interference, a simple relationship widely attributed to K. Mather follows. It states that among meioses in which one or more chiasmata occur in a given interval along the four-strand bundle, on average half of the resulting chromosomes will be recombinant across that interval. More formally, if r is the recombination fraction between two loci, and c_0 is the chance of having no chiasma located in the interval in any meiosis, then

$$r = \tfrac{1}{2}(1 - c_0). \qquad (4)$$

When $c_0 = c_0(d)$ depends only on the map length d of the interval, this relation is a map function. Now every chiasma involves just two of the four chromatids, and so the average number of chiasmata between two loci on the four-strand bundle is twice the average number of points of exchange between the same two loci on a single chromosome resulting from meiosis. Suppose that the number of chiasmata occurring in an interval along the four-strand bundle is Poisson distributed with mean $2d$. Then the map length of that interval is just d, and the chance of no chiasmata is e^{-2d}. Substituting into the above formula, we recover the Haldane map function (3) once more. It should be pointed out, however, that we can also recover this map function using a different distribution for the number of chiasmata and a different assumption concerning strand choice [13].

Keeping to the no chromatid interference assumption, we can derive many probabilistic models for recombination by postulating that chiasmata occur along the four-strand bundle according to a stationary renewal process (SRP). If the interchiasma density is f with respect to twice the map length density, then simple arguments from renewal theory show that for such models,

$$c_0(d) = 2\int_d^\infty \int_y^\infty f(t)\,\mathrm{d}t\,\mathrm{d}y. \tag{5}$$

It is shown in [14] that most of the map functions in the literature can be realized by substituting this expression with a suitable f into Mather's formula (4). This includes certain empirical map functions, such as the following suggested by Haldane in 1919,

$$M^{-1}(r) = 0.7r - 0.15\log(1 - 2r). \tag{6}$$

Map functions must satisfy certain constraints as a result of their definition, see [11] for details. Some functions suggested in the literature as suitable map functions do not satisfy these constraints [12], and should probably not be used. More importantly, most map functions are associated with stationary renewal processes whose multilocus recombination probabilities are extremely difficult to calculate, and for this reason are not so useful. The class of SRPs with **chi-square distributed** interchiasma distances in the map distance metric has proved both tractable and fairly general [14]. Another family of recombination models in the literature are termed the count-location processes [7]. These require the specification of a distribution $(g_k : k \geq 0)$ for the number (count) of chiasmata along the four-strand bundle, and a sequence F_k of functions giving the distribution of the locations of k chiasmata, given that k occur, $k \geq 1$. In the special case that F_k is equivalent to specifying k locations independently and identically according to a fixed distribution F, and no chromatid interference, we easily find that

$$c_0(d) = g\left(1 - \frac{2d}{m}\right), \tag{7}$$

where $g(s) = \sum_k g_k s^k$, $(0 < s < 1)$ and $m = g'(0)$. Risch & Lange [10] found that this class of recombination models did not give a very good fit to certain large data sets involving *Drosophila melanogaster*.

For many people, map functions are related to the notion of interference. Crossover interference is said to exist when the chance of one or more exchange points in an interval depends on the occurrence of exchange points in other, disjoint intervals. When the points of exchange form a Poisson process, there is no crossover interference. In general, such interference is observed, which is another reason why Poisson processes do not form suitable general models for recombination. (Note that a similar definition of interference can be formulated that refers to chiasmata occurring along the four-strand bundle. The corresponding notion is termed chiasma interference.) The traditional measure of interference is the coincidence coefficient, this being, for adjacent intervals A and B,

$$C_{A,B} = \frac{r(A\&B)}{r(A)r(B)}, \tag{8}$$

where $r(A)$ and $r(B)$ are the recombination fractions of A and B, respectively, and $r(A\&B)$ denotes the chance of simultaneous recombination across A and B. It is easy to check that

$$r(A\&B) = \frac{r(A) + r(B) - r(A \cup B)}{2},$$

where $A \cup B$ is the union of the adjacent intervals A and B. Suppose now that A has map length d, while B has small map length h, and that we take a limit (assumed to exist) in the expression for $C_{A,B}$ as $h \to 0$. Assuming that $M'(0) = 1$, which is one of the conditions that a map function must satisfy, we obtain the differential equation

$$M'(d) = 1 - 2C(d)M(d), \tag{9}$$

where $C(d)$ is the limiting coincidence coefficient, assumed to depend only on the map length of A.

This argument is due to Haldane [6], and many familiar map functions are solutions of this equation when $C(d)$ has the form $(M(d))^{n-1}$. For example, when $n = 1$, we get the Kosambi map function widely used in human genetics:

$$M(d) = \tfrac{1}{2}\tanh(2d). \tag{10}$$

The foregoing discussion shows that there is a connection between map functions and one aspect of crossover interference. In fact, this connection is quite superficial. A more useful (and outside of genetics more widely used) measure of interference is the expression $C_4(d) = C_{A,B}$ where A and B are infinitesimal intervals separated by a map distance d. This measure cannot, in general, be expressed in terms of the map function. In fact, there exist distinct probability models for recombination having the same map function, with one model having $C_4(d) =$ constant, while the other has $C_4(d)$, a function increasing almost monotonically from 0 at $d = 0$ to 1 for large d. In short, the two recombination models have the same map function, but very different interference properties, using the term interference in a general sense. Map functions do not adequately account for interference; this must be done using a probability model for recombination.

We close with some summary remarks. Map functions can be used to convert recombination fractions to map distances, correcting for multiple exchanges. They also correct for the effect of interference, but do not describe interference completely. They are essentially organism-dependent, and at best provide only rough approximations. It is not uncommon to see multilocus analyses carried out using the Poisson (no chiasma or crossover interference) model underlying Haldane's map function, at the end of the analysis correcting the estimated recombination fractions using Kosambi's or some other map function. This is necessary because map functions (such as Kosambi's) do not, in general, determine joint recombination probabilities for more than three loci. It is reassuring that this somewhat illogical approach gives estimated map distances that are not too different from those that would be obtained using (for example) a comparable stationary renewal process model with chi-square distributed interpoint distances. Ideally, multilocus mapping and linkage analyses should be carried out using a properly specified probability model for recombination suitable for the organism in question. When this is done, map functions are not needed.

References

[1] Bailey, N.T.J. (1961). *Introduction to the Mathematical Theory of Genetic Linkage*. Oxford University Press, London.

[2] Barratt, R.W., Newmeyer, D., Perkins, D.D. & Garnjobst, L. (1954). Map construction in Neurospora crassa, *Advances in Genetics* **6**, 1–93.

[3] Bridges, C.B. & Morgan, T.J. (1923). The second chromosome group of mutant characters, *Carnegie Institute of Washington Publication* **278**, Part II, 126–304.

[4] Fisher, R.A., Lyon, M.F. & Owen, A.R.G. (1947). The sex chromosome of the house mouse, *Heredity* **1**, 335–365.

[5] Goldgar, D.E. & Fain, P.R. (1988). Models of multilocus recombination: non-randomness in chiasma number and crossover location, *American Journal of Human Genetics* **43**, 38–45.

[6] Haldane, J.B.S. (1919). The combination of linkage values, and the calculation of distances between the loci of linked factors, *Journal of Genetics* **8**, 299–309.

[7] Karlin, S. (1984). Theoretical aspects of genetic map functions in recombination processes, in *Human Population Genetics: The Pittsburgh Symposium*, A. Chakravarti, ed. Van Nostrand Reinhold, New York, pp. 209–228.

[8] Nelson, S.F., McCusker, J.H., Sander, M.A., Kee, Y., Modrich, P. & Brown, P.O. (1993). Genomic mismatch scanning: A new approach to genetic linkage mapping, *Nature Genetics* **4**, 11–18.

[9] Ott, J. (1991). *Analysis of Human Genetic Linkage Data*. Johns Hopkins University Press, Baltimore.

[10] Risch, N. & Lange, K. (1979). An alternative model of recombination and interference, *Annals of Human Genetics* **43**, 61–70.

[11] Speed, T.P. (1996). What is a genetic map function?, in *Genetic Mapping and DNA sequencing*, T. Speed & M.S. Waterman, eds. Springer-Verlag, New York.

[12] Weeks, D.E. (1994). Invalidity of the Rao map function for three loci, *Human Heredity* **44**, 178–180.

[13] Zhao, H. (1995). Statistical analysis of genetical interference. PhD thesis, University of California at Berkeley.

[14] Zhao, H. & Speed, T.P. (1996). On genetic map functions, *Genetics* **142**, 1369–1377.

T.P. SPEED