Point and Interval Estimates of Marker Location in Radiation Hybrid Mapping

Heather M. Stringham,[1] Michael Boehnke,[1] and Kenneth Lange[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor; and [2]Departments of Biomathematics and Human Genetics, University of California, Los Angeles

Summary

Radiation hybrid (RH) mapping is a powerful method for ordering loci on chromosomes and for estimating the distances between them. RH mapping is currently used to construct both framework maps, in which all markers are ordered with high confidence (e.g., 1,000:1 relative maximum likelihood), and comprehensive maps, which include markers with less-confident placement. To deal with uncertainty in the order and location of markers, marker positions may be estimated conditional on the most likely marker order, plausible intervals for nonframework markers may be indicated on a framework map, or bins of markers may be constructed. We propose a statistical method for estimating marker position that combines information from all plausible marker orders, gives a measure of uncertainty in location for each marker, and provides an alternative to the current practice of binning. Assuming that the prior distribution for the retention probabilities is uniform and that the marker loci are distributed independently and uniformly on an interval of specified length, we calculate the posterior distribution of marker position for each marker. The median or mean of this distribution provides a point estimate of marker location. An interval estimate of marker location may be constructed either by using the 100(α/2) and 100(1 − α)/2 percentiles of the distribution to form a 100(1 − α)% posterior credible interval or by calculating the shortest 100(1 − α)% posterior credible interval. These point and interval estimates take into account ordering uncertainty and do not depend on the assumption of a particular marker order. We evaluate the performance of the estimates on the basis of results from simulated data and illustrate the method with two examples.

Received December 21, 1998; accepted for publication June 15, 1999; electronically published July 12, 1999.

Address for correspondence and reprints: Dr. Michael Boehnke, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

0002-9297/99/6502-0031$02.00

Introduction

Radiation hybrid (RH) mapping (Goss and Harris 1975; Cox et al. 1990; Walter et al. 1994) is a powerful method for ordering loci on mammalian chromosomes and for estimating the physical distances between them. Several statistical methods have been developed for the analysis of RH mapping data. These include methods based on maximum likelihood and minimum obligate chromosome breaks (e.g., see Boehnke et al. 1991; Lange et al. 1995). Programs implementing these methods are useful for ordering modest numbers of loci or for constructing framework maps from large numbers of loci. In addition, heuristic methods based on simulated annealing (D. R. Cox, personal communication), graph theory and greedy algorithms (Slonim et al. 1997), minimizing the sum of the distances between adjacent markers (Xia 1997), and artificial intelligence paradigms (Matise and Chakravarti 1995) have been developed that efficiently order large numbers of loci.

Common to all of these methods is the problem of uncertainty in order and location for some groups of markers. Although framework maps consisting of a subset of markers that can be ordered with a strong level of support (e.g., 1,000:1 maximum likelihood ratio) can generally be constructed, attempts to include nonframework markers in the map often involve a degree of uncertainty that should be acknowledged and dealt with. Often, the "best" order among several plausible ones is chosen, and map distances are estimated conditional on that order. Unfortunately, there is not always a clearly best order, making it difficult and/or undesirable to choose a single order on which to condition. As an alternative, all plausible intervals for each nonframework marker may be indicated on a framework map, along with the relative odds for each of these intervals. This alternative deals honestly with the uncertainty in marker ordering but gives no estimate of location and no indication of the plausibility of different orders for markers falling within the same interval of the framework map. Binning of nonframework markers is also a common practice, with the same disadvantage of giving no indication of the most likely orders or positions for markers within a particular bin.

In this article, we propose a Bayesian method of mapping groups of markers that takes into account all plausible locus orders and that provides point estimates of location and interval estimates that represent the uncertainty of location for each marker. We evaluate the performance of these estimates on the basis of results from simulated data and illustrate the method with two examples.

Methods

Notation and Assumptions

We estimate the marker positions, within a map, relative to that of a particular marker that we choose to be the "anchor" of the map. In particular, we calculate for each marker the posterior distribution of the distance of that marker from the anchor. To calculate this distribution, we require several assumptions. We assume as our prior distribution that m marker loci are independently and uniformly distributed on an interval of specified length D. We further assume that all chromosome fragments generated by irradiating the mammalian donor cell have the same retention probability r, which, for the sake of simplicity, is postulated to have a uniform prior distribution on the interval (0,1). In our analysis, we focus on the interlocus distances, d_i, where d_i is the distance between adjacent markers i and $i+1$ (see fig. 1). Here, d_i is measured in Rays, where 1 Ray corresponds to one expected break per hybrid. We assume that the chromosome breakage occurs at random along the chromosome and so follows a Poisson process; consequently, we use the Haldane mapping function $d_i = -\ln(1-\theta_i)$ relating d_i to the probability θ_i of at least one break between markers i and $i+1$.

Conditional on a marker order γ and particular values for r and the vector of interlocus distances $\mathbf{d} = (d_1, d_2, \ldots, d_{m-1})$, we can calculate the probability (or likelihood) of the observed RH mapping data, using the theory of hidden Markov chains. The log likelihood

$$\mathrm{L}_\gamma = \log \sum_{g_1} \cdots \sum_{g_m} \binom{c}{g_1} r^{g_1}(1-r)^{c-g_1} \prod_{i=1}^{m-1} t_{c,i}(g_i, g_{i+1}) \ ,$$

where g_i is the state of the Markov chain at locus i (i.e., the number of copies of locus i that are retained) and $t_{c,i}(j,k)$ is the probability, for a c-ploid hybrid, of a transition from state j at locus i to state k at locus $i+1$. The log prior $\mathrm{R}_\gamma = \log \frac{m!}{\mathrm{D}^m}(\mathrm{D} - d_1 - \cdots - d_{m-1})$ for $\{d_i : \Sigma_{i=1}^{m-1} d_i \leqslant \mathrm{D}\}$. The log posterior is proportional to the sum $\mathrm{L}_\gamma + \mathrm{R}_\gamma$. For details, see the work of Lange et al. (1995) and Lange (1997).

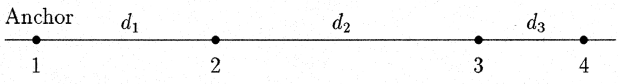

Figure 1 Hypothetical map of four loci with marker locus 1 as the anchor. d_1, d_2, and d_3 are interlocus distances, measured in Rays.

Posterior Distribution of Marker Position Conditional on Marker Order

Given a specific marker order γ, we assume that the posterior density of the parameter vector $\phi = (\mathbf{d}, r)$ is approximately multivariate normal with mean vector $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$. Here $\hat{\mu}$ is the posterior mode, found by maximizing the sum of the log likelihood L_γ and the log prior R_γ, and $\hat{\Sigma} = (\hat{\sigma}_{ij}) = -\hat{\Omega}^{-1}$, where $\hat{\Omega}$ is the matrix of second partial derivatives of the log posterior, evaluated at the posterior mode. Under this assumption, the conditional posterior distribution of d_i is normal with mean $\hat{\mu}_i$ and variance $\hat{\sigma}_{ii}$. Consequently, the conditional posterior distribution of the distance $d_i + \ldots + d_{j-1}$ between loci i and j is normal with mean $\Sigma_{k=i}^{j-1}\hat{\mu}_k$ and variance $\Sigma_{k=i}^{j-1}\Sigma_{l=i}^{j-1}\hat{\sigma}_{kl}$. Thus, if we pick an anchor marker from which to calculate distances, we can calculate the univariate posterior distribution F_γ of a particular marker position relative to the anchor, conditional on an order γ. For example, if marker 1 is the anchor, then the conditional distribution F_γ of the distance from the anchor to marker 4, for order $\gamma = (1,2,3,4)$, is normal with mean $\mu = \hat{\mu}_1 + \hat{\mu}_2 + \hat{\mu}_3$ and variance $\sigma^2 = \Sigma_{i=1}^{3}\Sigma_{j=1}^{3}\hat{\sigma}_{ij}$.

Unconditional Posterior Distribution of Marker Position

Given the parameter vector $\phi = (\mathbf{d}, r)$ and our implicit assumption that all locus orders have equal prior probabilities, the posterior probability P_γ of marker order γ can be calculated, by Bayes's theorem, as

$$P_\gamma \quad = \quad \frac{\int e^{\mathrm{L}_\gamma(\phi)+\mathrm{R}_\gamma(\phi)} d\phi}{\sum_\nu \int e^{\mathrm{L}_\nu(\phi)+\mathrm{R}_\nu(\phi)} d\phi} \ , \qquad (1)$$

where ν ranges over all $m!/2$ possible marker orders. This expression represents the proportion of the posterior likelihood, averaged over the parameter space, that is attributable to order γ. The integrals in equation (1) can be evaluated by using a Laplace approximation (de Bruijn 1981; Tierney and Kadane 1986), in which the log posterior is expanded in a second-order Taylor series about the posterior mode $\hat{\mu}$. This yields

$$\int e^{\mathrm{L}(\phi)+\mathrm{R}(\phi)}d\phi \approx e^{\mathrm{L}(\hat{\mu})+\mathrm{R}(\hat{\mu})}(2\pi)^{m/2}\det(-\hat{\Omega})^{-1/2} \ ,$$

where det represents matrix determinant. When the number of possible orders $m!/2$ is large, summing over all possible orders is impractical. Instead, the sum in the denominator of equation (1) can be approximated by including only the most plausible orders ν (Lange et al. 1995; Lange 1997).

With this posterior probability P_γ of marker order γ in hand, the unconditional posterior distribution of distance from the anchor can then be calculated as the mixture of normal distributions $\Sigma_\gamma P_\gamma \mathrm{F}_\gamma$, where in principle γ ranges over all possible orders but, in practice, will again be restricted to a subset of orders. A procedure for selecting the set of plausible orders over which to sum is described below (see the Implementation Issues section).

Estimates of Marker Position

As a point estimate for marker position, we use the median x of the posterior distribution of marker position. To obtain this estimate, we numerically solve the equation $\Sigma_\gamma P_\gamma \mathrm{F}_\gamma(x) = .50$, where the sum is over the set of plausible orders. If μ_γ is the mean of F_γ, then the posterior mean $\mu. = \Sigma_\gamma P_\gamma \mu_\gamma$ can also be used as a point estimate for marker position.

To construct interval estimates for marker position, we use the $100(\alpha/2)$ and $100(1-\alpha)/2$ percentiles of the posterior distribution to form a $100(1-\alpha)\%$ posterior credible interval for the position of a marker. For example, a 90% posterior credible interval (x_l,x_r) can be found by solving numerically the equations $\Sigma_\gamma P_\gamma \mathrm{F}_\gamma(x_l) = .05$ and $\Sigma_\gamma P_\gamma \mathrm{F}_\gamma(x_r) = .95$. Alternatively, the shortest $100(1-\alpha)\%$ posterior credible interval can be computed. We compute this shortest interval by selecting the smallest interval from a grid search, refining the left or right endpoint by using 10 steps of Newton's method (Lange 1997), and adjusting the opposite endpoint accordingly. These estimates can be calculated for each marker except the anchor, to produce a map such as that shown in figure 2, giving a point estimate for the position of each marker and a posterior credible interval for each marker relative to the anchor.

Implementation Issues

Three issues arise in the implementation of our method for marker positioning: how to determine the set of plausible orders to consider, how to choose the anchor marker, and how to choose the map interval length for the prior distribution of marker positions. At first glance, the question of determining the set of plausible orders to consider does not seem to be a difficult one. It appears that one could simply name an acceptable cutoff for the relative likelihood of an order and use all orders with greater likelihood than that cutoff point. The issue is more complicated than that, however, for we must, in addition, decide the orientation of each marker order along the chromosome. It is also important for the markers to be consistently oriented with respect to the anchor. One solution to this problem would be to use the absolute value of the distance between the markers. However, if we did that, all distance estimates would be positive, and we would lose our ability to determine which markers are on the same side of the anchor and which are separated by the anchor. A better solution is to orient each marker order so that three loci are consistently oriented in each order that we choose to consider.

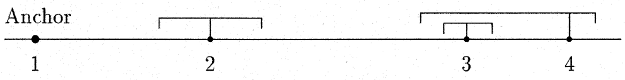

Figure 2 Hypothetical map with three loci mapped relative to anchor marker 1. Point estimates are indicated by a vertical line and dot; interval estimates are indicated by a horizontal line with short vertical lines denoting the interval endpoints.

To determine the set of plausible orders and the anchor, we first generate a list of marker orders ranked by maximum likelihood. We then consider all marker triples (i,j,k) from the set of m markers. Each triple is a subset of size three, with a designated middle marker. There are $m(m-1)(m-2)/2$ such triples, if we neglect orientation. We count the number of marker orders, consecutive from the most likely to the least likely, in which the triple is consistently ordered. The middle marker of the best triple provides the anchor, and the orders consistent with this triple constitute the set of plausible orders, which can then be oriented consistently. As an example, consider the six markers in table 4A. The triple of markers (1,2,3) is ordered consistently in the two most likely orders; triples (1,2,4), (1,2,5), and (1,2,6) are ordered consistently in the first three orders; triples (1,3,4), (1,3,5), and (1,3,6) are ordered consistently in four orders; and triples (1,4,5), (1,4,6), (2,4,5), (2,4,6), (3,4,5), and (3,4,6) are ordered consistently in the five best orders. No triple could be consistently ordered in more than the five most likely orders given in the table (data not shown). We chose marker 4 as anchor because it is the middle locus of all triples ordered consistently in the five most likely orders. If there are several possible choices for the anchor marker, one might wish to choose the triple with middle locus nearest the center of the map or in a region of particular interest.

Given *m* points placed at random on an interval of length D, the expected distance between the first and last points is $\frac{m-1}{m+1}$D. Thus, a reasonable choice of map interval length for the prior distribution of marker positions is $\frac{m+1}{m-1}\hat{D}$, where $\hat{D}$ is the maximum likelihood estimate of map length. Lange et al. (1995) have noted that this choice of interval length may be too confining in practice and suggest that it be inflated by 10%–20%. We explore the impact of the choice of prior map length in our simulations.

Simulations

To test the accuracy and usefulness of our methods, we simulated RH mapping data under several maps and applied the method to each data set, comparing our results with the true maps. For each simulated data set, we generated data for 10 markers typed on 100 diploid RHs with equal retention probability $r = .10$ per chromosome or, equivalently, $1 - (1 - .10)^2 = .19$ per hybrid. We generated data, assuming maps with equally spaced markers at 30 and 15 cR and with randomly spaced markers with average spacings of 30 and 15 cR. We followed the procedure described above for determining the set of plausible orders and the anchor. When several triples tied for the greatest number of consistent orders, we preferentially selected the triple in which the middle marker was near the middle of the map. We compared the properties of this approach to selecting preferentially the triple with middle locus nearest the end of the map.

We analyzed the simulated RH data by using prior map lengths ranging from $\frac{m+1}{m-1}\hat{D}$ to $1.5 * \frac{m+1}{m-1}\hat{D}$ and compared the resulting estimated map lengths with the true map length $D = \Sigma_{d=1}^{m-1} d_i$ and the length $\hat{D}$ of the best maximum likelihood map. To determine the best point and interval estimators, we compared the median and the mean of the posterior distribution of marker position and compared two types of 90% interval estimates: the 5th–95th percentiles of the posterior distribution and the shortest 90% posterior credible interval. Finally, we compared the success of the method in determining the correct order of the markers to that of using the best maximum likelihood order.

Results

Simulations

Table 1 gives the results of our simulations. We use the distance between point estimates for the leftmost and rightmost markers as an estimate of map length. The bias in estimated map length is nearest 0 for a prior map length of $1.3 * \frac{11}{9}\hat{D}$ (or $1.5 * \frac{11}{9}\hat{D}$) for markers spaced 30 cR (or 15 cR) apart (table 1, cols. 1 and 2). It tends to be smaller when the median (M_1) is used as the point estimate in calculating the estimated map length, particularly for randomly spaced markers. The mean squared error (MSE) of the estimated map length (table 1, cols. 3 and 4) shows no consistent pattern as a function of prior map length but was smaller, in all cases, than the MSE of the best maximum likelihood map length.

The biases in the median and mean point estimates of marker position are similar and tend to decrease with increasing prior map length (table 1, cols. 5 and 6). The MSE of the point estimates is smaller for the mean (M_2) than for the median and tends to increase with increasing prior map length (table 1, cols. 7 and 8). The median and mean estimates of marker position behave similarly in their ability to order the markers (table 1, cols. 9 and 10). Both order the markers correctly with a frequency similar to that of maximum likelihood. Probability of correct ordering is greater for equally spaced markers and for markers spaced 30 cR apart.

Coverage of the true marker position by both types of interval estimators is less than the nominal 90% but increases with prior map interval length (table 1, cols. 11 and 12). This increase in coverage probability is most likely due to increasing interval width with prior map length. Coverage is better for markers spaced 30 cR apart than for those spaced 15 cR apart and is better for equally spaced markers than for unequally spaced markers. Intervals constructed by using the 5th and 95th percentiles perform similarly to the shortest 90% credible intervals.

Coverage is also affected by anchor choice. Interval coverage probabilities are slightly lower when anchors are chosen near the end of the map, even though interval widths are increased. In addition, the bias and MSE of the point estimates are greater when these anchors are used. Choice of anchor has little effect on the ability of the method to order markers correctly or on the bias or MSE of the estimated map length (data not shown).

Applications

To illustrate our method, we present two examples. The first uses 13 sequence-tagged site (STS) markers on chromosome 4p that were typed on 83 RHs from the Stanford G3 panel distributed by Research Genetics (Lange et al. 1995). Note that we exclude the first of the 14 markers in the original data because it appears to fall into a separate linkage group. Markers were considered to fall into the same linkage group if the maximum pairwise LOD score was >6.0 between each marker and at least one other in the group. Table 2 lists the 17 most likely orders for the 13 markers, as determined by maximum likelihood. The three rightmost columns of the table give the relative maximum likelihoods and two posterior probabilities based on prior map

Table 1

Performance of Point and Interval Estimates of Marker Location

	Estimated Map Length				Marker Position							
	Mean Bias		MSE		Mean Bias		MSE		Correct Order (%)		Coverage (%)	
Prior Map Length	M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2	M_1	M_2	Shortest	5–95
	Equally Spaced Markers, 30 cR; ML Order Correct in 96.9%; ML Map Bias −3.6, MSE 1,661											
$1.0 * \frac{11}{9}\hat{D}$	3.4	3.5	1,458	1,432	1.8	1.8	387	380	96.7	96.9	86.7	86.7
$1.1 * \frac{11}{9}\hat{D}$	1.7	1.8	1,436	1,402	1.7	1.7	398	389	96.7	97.0	88.0	88.0
$1.2 * \frac{11}{9}\hat{D}$	.6	.8	1,443	1,403	1.6	1.7	407	397	96.7	97.0	88.7	88.7
$1.3 * \frac{11}{9}\hat{D}$	−.1	.1	1,457	1,411	1.6	1.6	415	404	96.7	96.9	89.2	89.2
	Equally Spaced Markers, 15 cR; ML Order Correct in 88.7%; ML Map Bias −4.3, MSE 726											
$1.0 * \frac{11}{9}\hat{D}$	5.8	5.8	653	653	.8	.8	161	161	88.6	88.5	84.4	84.4
$1.1 * \frac{11}{9}\hat{D}$	3.6	3.6	639	639	.7	.7	164	164	88.6	88.6	86.1	86.1
$1.2 * \frac{11}{9}\hat{D}$	2.1	2.1	639	639	.6	.6	168	168	88.7	88.6	87.1	87.1
$1.3 * \frac{11}{9}\hat{D}$	1.1	1.1	643	643	.6	.6	171	171	88.6	88.6	87.8	87.8
$1.4 * \frac{11}{9}\hat{D}$	.4	.4	648	648	.5	.5	173	173	88.7	88.5	88.3	88.3
$1.5 * \frac{11}{9}\hat{D}$	−.2	−.2	654	653	.5	.5	175	176	88.7	88.5	88.6	88.6
	Randomly Spaced Markers, Average 30 cR; ML Order Correct in 52.6%; ML Map Bias −5.1, MSE 4,050											
$1.0 * \frac{11}{9}\hat{D}$	6.4	11.9	2,076	3,609	.1	.0	2,229	1,746	52.2	51.3	84.9	85.1
$1.1 * \frac{11}{9}\hat{D}$	4.8	11.1	1,972	3,667	.1	−.1	2,203	1,755	52.2	51.0	86.5	86.7
$1.2 * \frac{11}{9}\hat{D}$	3.8	10.9	1,921	3,779	.1	−.2	2,199	1,780	52.2	50.8	87.7	87.8
$1.3 * \frac{11}{9}\hat{D}$	3.2	10.9	1,892	3,918	.1	−.3	2,192	1,814	52.3	50.8	88.4	88.5
	Randomly Spaced Markers, Average 15 cR; ML Order Correct in 37.7%; ML Map Bias −3.7, MSE 1,210											
$1.0 * \frac{11}{9}\hat{D}$	7.7	8.7	787	863	.2	.3	286	256	37.0	37.1	81.4	81.4
$1.1 * \frac{11}{9}\hat{D}$	5.5	6.6	761	833	.1	.2	289	257	37.1	37.1	83.2	83.3
$1.2 * \frac{11}{9}\hat{D}$	3.9	5.1	753	822	.1	.2	292	259	37.1	37.1	84.5	84.6
$1.3 * \frac{11}{9}\hat{D}$	2.8	4.1	752	818	.1	.1	296	261	37.0	37.1	85.2	85.3
$1.4 * \frac{11}{9}\hat{D}$	2.0	3.3	753	817	.1	.1	299	262	37.0	37.1	86.0	86.1
$1.5 * \frac{11}{9}\hat{D}$	1.4	2.8	756	817	.1	.1	301	264	37.0	37.1	86.4	86.5

NOTE.—Results based on 1,000 simulated data sets. Point estimates: M_1 = median; M_2 = mean. Bias in map length: $100 * (D - D_i)/D$ where D = true distance in cR between first and last markers and D_i uses point estimates M_i for positions of first and last markers. Bias in point estimates: true marker position $-M_i$ (cR) (i = 1,2). Interval estimates: shortest = shortest 90% posterior credible interval; 5–95 = 90% posterior credible interval constructed using the 5th and 95th percentiles. ML = maximum likelihood.

lengths of $1.3 * \frac{14}{12}\hat{D} = 387$ cR and $1.5 * \frac{14}{12}\hat{D} = 446$ cR. Figure 3*a* shows the maximum likelihood map, figure 3*b* and *c* shows the maps generated by our method for two choices of anchor, and figure 3*b* and *d* shows the effect of using two different prior map lengths. In each case, we used the median of the posterior distribution as our point estimate of marker position and used the shortest 90% posterior credible interval as our interval estimate. We chose marker 8 as anchor because it is the middle of three consistently ordered loci (7,8,10) found by the procedure described above. Marker 10 is the end locus of this triple and the middle locus of an equivalent triple (8,10,13). Notice that all maps are similar to the maximum likelihood map and that anchor choice does not greatly affect the point estimates. Interval estimates tend to become wider for markers farther from the anchor. Choice of prior map length does not have a large effect, although the lengths of the credible intervals are increased slightly for larger prior map length (fig. 3*d*).

The second example includes 32 STS markers on chromosome 1q that were typed on 83 RHs from the Stanford G3 panel. These markers include all those placed by the Stanford Human Genome Center into bins 97–114, plus two flanking markers. These 32 markers fall into three linkage groups. We analyzed each linkage group separately. We broke the second linkage group into two subgroups after initial analyses, by maximum likelihood and our method, suggested a gap of 45–50 cR between the first six and the last seven markers. We chose anchors by selecting the middle locus of the marker triple(s) ordered consistently in the greatest number of most likely marker orders. For the third linkage group, markers 4 and 5 both met this criterion, and we arbitrarily selected marker 5 as anchor. We again used the median and the shortest 90% interval as our estimates. Tables 3–5 and figure 4 present our results. Although the marker order within some bins (e.g., bin 109) remains ambiguous, we were able to order many of the markers. The markers within bins 102, 103, 104, and 107 are unambiguously ordered. Although some ambiguity remains in bin 108, our map clearly improves on binning. Our results agree with the 9-marker 1,000:1 and 22-marker 100:1 framework maps constructed, by maximum likelihood, from the full set of 32 markers (results not shown).

Discussion

The method that we have described provides estimates of marker position that take into account ordering uncertainty and that are not conditional on a particular marker order. It can be used to refine the order within bins of markers produced by other methods. It successfully combines information from all plausible orders, and the posterior credible intervals that it generates provide a useful measure of the uncertainty in marker positions.

We have given some practical solutions to the implementation issues regarding the appropriate set of plausible orders, anchor choice, and prior map length. The question of how many plausible orders to consider is an important one. Since implementation of our method requires three markers to be consistently oriented in all orders considered, it is not always possible to use all orders that appear fairly plausible by the maximum likelihood criterion. This will often be true with small sets of markers (e.g., 4 or 5). In each of our examples here, however, we were able to use all orders within at least

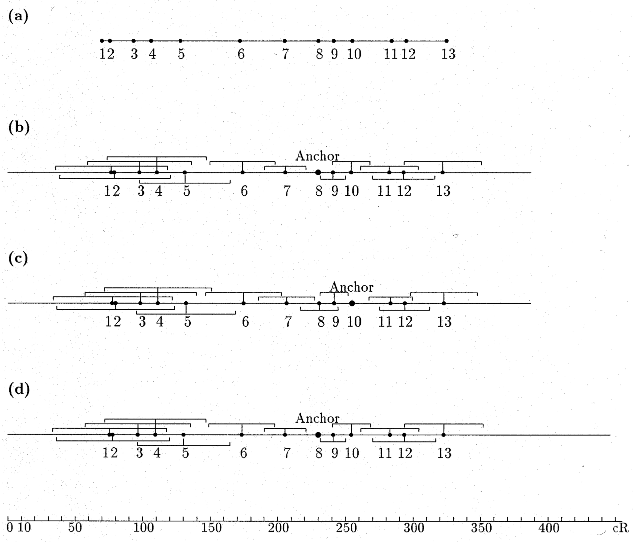

Figure 3 Maps of 13 markers on chromosome 4p. *a,* maximum likelihood map of the best order ($\hat{D}$ = 255 cR). *b,* Prior map length 1.3 ∗ $\frac{14}{12}\hat{D}$ = 387 cR, anchor = 8. *c,* Prior map length 387 cR, anchor = 10. *d,*Prior map length 1.5 ∗ $\frac{14}{12}\hat{D}$ = 446 cR, anchor = 8.

Table 2

Most Likely Locus Orders for 13 Markers on Chromosome 4p

														P_γ[c]	
						γ[a]							$\Delta\log_{10}L$[b]	387 cR	446 cR
1	2	3	4	5	6	7	8	9	10	11	12	13	.0000	.70008	.70078
2	*1*	3	4	5	6	7	8	9	10	11	12	13	.3598	.29735	.29646
5	*4*	*3*	*2*	*1*	6	7	8	9	10	11	12	13	2.7331	.00127	.00135
1	2	3	4	5	6	7	8	9	10	*12*	*11*	13	3.0950	.00063	.00066
2	*1*	3	4	5	6	7	8	9	10	*12*	*11*	13	3.4509	.00027	.00028
6	*5*	*4*	*3*	*1*	*2*	7	8	9	10	11	12	13	3.7457	.00013	.00014
5	*4*	*3*	*1*	*2*	6	7	8	9	10	11	12	13	3.7860	.00011	.00013
6	7	8	9	10	11	12	13	*5*	*4*	*3*	*2*	*1*	4.3378	.00003	.00004
1	2	3	4	5	*7*	*6*	8	9	10	11	12	13	4.4017	.00003	.00004
6	1	2	3	4	*5*	7	8	9	10	11	12	13	4.6905	.00002	.00002
6	*5*	*4*	*3*	*2*	*1*	7	8	9	10	11	12	13	4.7126	.00001	.00002
6	7	8	10	9	11	12	13	*5*	*4*	*3*	*1*	*2*	4.7207	.00001	.00002
2	*1*	3	4	5	*7*	*6*	8	9	10	11	12	13	4.7643	.00001	.00002
6	7	8	9	10	11	12	13	*1*	*2*	*3*	*4*	*5*	4.7705	.00001	.00002
6	7	8	9	10	11	12	13	*2*	*1*	*3*	*4*	*5*	4.9425	.00001	.00001
1	2	3	4	5	6	7	8	9	10	*13*	*12*	*11*	5.1247	.00001	.00001
1	2	3	4	5	6	7	*9*	*8*	10	11	12	13	5.3123	.00000	.00001

[a] Italics indicate rearrangements relative to the most likely order.

[b] $\Delta\log_{10}L$ indicates the difference in maximum log likelihood from the best order.

[c] P_γ is the posterior probability of order γ for the prior map length indicated. Orders with $P_\gamma < .000005$ for both map lengths are not shown.

Table 3

Most Likely Orders for Chromosome 1q Markers in Linkage Group 1

					γ^{a}						$\Delta\log_{10}L^{b}$	P_{γ}^{c}
1	2	3	4	5	6	7	8	9	10	11	.0000	.90490
1	2	3	4	5	6	7	8	*11*	*10*	*9*	1.2645	.05467
1	2	3	4	5	6	7	8	*10*	*11*	*9*	1.6888	.02434
1	2	3	4	5	6	7	8	*10*	*9*	11	2.3679	.00492
1	2	3	4	5	6	7	8	9	*11*	*10*	2.4388	.00397
1	2	3	4	5	6	*8*	*7*	9	10	11	2.6338	.00293
1	2	3	4	5	*7*	*6*	8	9	10	11	2.9135	.00171
2	*1*	3	4	5	6	7	8	9	10	11	3.1505	.00090
1	2	3	4	*8*	*7*	*6*	*5*	*11*	*10*	*9*	3.6923	.00024
1	2	3	4	5	*8*	*7*	*6*	*11*	*10*	*9*	3.7424	.00022
1	2	3	4	5	6	*8*	*7*	*11*	*10*	*9*	3.7545	.00023
3	2	*1*	4	5	6	7	8	9	10	11	4.0858	.00007
1	2	3	4	*7*	*8*	*6*	*5*	*11*	*10*	*9*	4.1101	.00011
1	2	3	4	5	*7*	*6*	8	*11*	*10*	*9*	4.1777	.00010
1	2	3	4	5	*7*	*8*	*6*	*11*	*10*	*9*	4.1805	.00010
1	*3*	*2*	4	5	6	7	8	9	10	11	4.2564	.00006
1	2	3	4	*8*	*7*	*6*	*5*	9	10	11	4.3420	.00005
1	2	3	4	5	6	*8*	*7*	*10*	*11*	*9*	4.3445	.00007
1	2	3	4	5	*8*	*7*	*6*	9	10	11	4.3981	.00005
2	*1*	3	4	5	6	7	8	*11*	*10*	*9*	4.4068	.00005
1	2	3	4	5	6	7	8	*11*	*9*	*10*	4.4352	.00004
1	2	3	*5*	*4*	6	7	8	9	10	11	4.4409	.00004
1	2	3	4	5	*7*	*6*	8	*10*	*11*	*9*	4.6031	.00004
1	2	3	4	*7*	*8*	*6*	*5*	9	10	11	4.7594	.00002
2	*1*	3	4	5	6	7	8	*10*	*11*	*9*	4.8301	.00002
1	2	3	4	5	*7*	*8*	*6*	9	10	11	4.8352	.00002
1	2	3	*5*	*4*	*7*	*6*	8	9	10	11	4.8591	.00002
1	2	3	4	5	6	*8*	*7*	*10*	*9*	11	5.0214	.00001
11	*10*	*9*	1	2	3	4	5	6	7	8	5.0401	.00001
1	2	3	4	5	6	*8*	*7*	9	*11*	*10*	5.0722	.00001
1	2	3	4	5	*8*	*6*	*7*	9	10	11	5.1118	.00001
1	2	3	4	5	*11*	*10*	*9*	*8*	*7*	*6*	5.1655	.00001
3	1	2	4	5	6	7	8	9	10	11	5.2185	.00001
1	2	3	4	5	*7*	*6*	8	*10*	*9*	11	5.2811	.00001
3	2	*1*	4	5	6	7	8	*11*	*10*	*9*	5.3460	.00000
1	2	3	4	5	*7*	*6*	8	9	*11*	*10*	5.3518	.00001
1	2	3	4	*7*	*6*	*5*	8	9	10	11	5.4076	.00001
1	2	3	4	6	*5*	7	8	9	10	11	5.4279	.00001

[a] Italic indicates rearrangements relative to the most likely order.
[b] $\Delta\log_{10}L$ indicates the difference in maximum log likelihood from the best order.
[c] Orders with $P_{\gamma} < .000005$ are not shown.

1,000:1 odds of the best maximum likelihood order. The method may not be as accurate for data sets in which only a few of the plausible orders can be used.

Other procedures for determining the set of plausible orders are also possible. For example, an alternative procedure would be to generate a list of marker orders ranked by maximum likelihood, calculate the posterior probabilities of these orders, and rank them by their posterior probability rather than by their maximum likelihood, ignoring orders with very small posterior probability, say <.000005. Triples of loci could then be ranked, and an anchor marker could be chosen in the same manner as before. Results in most cases should be nearly the same as in the procedure that we have used.

Both of these procedures depend on being able to generate a list of orders ranked by maximum likelihood. This may not be possible if the number of orders to consider is extremely large. Breaking the region into several smaller groups of markers, as in our second example, can be helpful. It is also possible that generating a list ranked by another criterion would work in such cases.

The ability of the method to order markers correctly does not seem sensitive to the choice of the anchor or to the prior map length. However, these choices do affect the coverage probabilities of the intervals and the bias and MSE of the point estimates. There are often several suitable choices for the anchor. Map position tends to be measured with greater precision for markers near the anchor (data not shown), so it makes sense to choose

Table 4

Most Likely Orders for Chromosome 1q Markers in Linkage Groups 2a and 2b

γ^a							$\Delta\log_{10}L^b$	P_γ
A. Linkage Group 2a								
1	2	3	4	5	6		.0000	.99369
1	2	3	4	*6*	*5*		2.6034	.00264
1	*3*	*2*	4	5	6		2.7269	.00204
2	*1*	3	4	5	6		2.9196	.00155
3	2	*1*	4	5	6		4.0731	.00008
B. Linkage Group 2b								
7	8	9	10	11	12	13	.0000	.46003
7	8	9	10	*12*	*11*	13	.0002	.45997
9	*8*	*7*	10	11	12	13	1.4781	.01417
9	*8*	*7*	10	*12*	*11*	13	1.4784	.01417
8	9	10	*12*	*11*	13	*7*	1.5588	.01134
8	9	10	11	12	13	*7*	1.5594	.01133
8	*9*	*7*	10	11	12	13	1.6355	.00881
8	*9*	*7*	10	*12*	*11*	13	1.6359	.00880
7	*9*	*8*	10	11	12	13	2.2780	.00262
7	*9*	*8*	10	*12*	*11*	13	2.2783	.00262
7	8	9	10	*13*	*11*	*12*	2.2846	.00251
8	*7*	9	10	11	12	13	2.6074	.00110
8	*7*	9	10	*12*	*11*	13	2.6076	.00110
7	8	9	10	11	*13*	*12*	2.7134	.00115
7	8	9	10	*13*	*12*	*11*	3.2825	.00028

[a] Italics indicate rearrangements relative to the most likely order.

[b] $\Delta\log_{10}L$ indicates the difference in maximum log likelihood from the best order.

an anchor as close as possible to the region of greatest interest or, given no such region, the center of the map. Marker retention could also be considered when one is choosing among several candidates for the anchor.

Small or even moderate changes in the prior specification of map length (e.g., ±50%) do not generally result in much change in the point estimates of marker position. However, as one might anticipate, the average lengths of the interval estimates increase with increasing prior map length. In our simulations, increasing the prior map length from $\frac{m+1}{m-1}\hat{D}$ to $1.3 * \frac{m+1}{m-1}\hat{D}$ resulted in an average increase in interval length of 9.5% for the shortest 90% intervals (data not shown). Using prior map lengths from $1.3 * \frac{m+1}{m-1}\hat{D}$ to $1.5 * \frac{m+1}{m-1}\hat{D}$ resulted in the most accurate estimates of map length in our simulations. Although the map length was consistently overestimated by the best maximum likelihood map, we were able to get quite close to the true map length by using these prior map lengths. Because increased prior map lengths also tend to give more-accurate point estimates and interval coverage closer to the nominal, we recommend the use of prior map lengths from $1.3 * \frac{m+1}{m-1}\hat{D}$ to $1.5 * \frac{m+1}{m-1}\hat{D}$.

We have used the mean and the median of the posterior distribution as point estimates of marker position. In our experience, these estimates generally give similar results, and we have chosen to use the median. We have used the $100(\alpha/2)$ and $100(1-\alpha)/2$ percentiles to construct a $100(1-\alpha)\%$ posterior credible interval for marker location; we have also constructed the shortest $100(1-\alpha)\%$ credible interval. In our simulations, we have found these intervals to be similar, differing in length by <1 cR for the large majority of markers with average spacings of 30 or 15 cR (data not shown).

Note that overlapping intervals (such as those for markers 3 and 4 in fig. 2) do not necessarily imply that those markers are not well ordered relative to one another, since the distributions for marker position are calculated only with respect to the anchor. Markers that are not well ordered more often result in point estimates that nearly coincide (e.g., see markers 11 and 12 in fig. 4*b* and markers 1 and 2 in fig. 4*c*).

Our method provides an effective way to use information from all plausible orders, not just the most likely order. The graphics based on the method make it easy to see which marker locations are most precise (in terms of width of posterior credible intervals) and can be a useful visual tool for summarizing the best maximum likelihood orders. For example, consider the 32 markers in our second example. Our method places these markers in the same order as the best maximum likelihood order for the full set of 32 markers. The second most likely

Table 5

Most Likely Orders for Chromosome 1q Markers in Linkage Group 3

γ^a								$\Delta\log_{10}L^b$	P_γ
2	1	3	4	5	6	7	8	.0000	.41328
1	*2*	3	4	5	6	7	8	.0004	.41286
2	1	3	4	5	6	*8*	7	.7813	.08228
1	*2*	3	4	5	6	*8*	7	.7813	.08227
2	1	3	4	5	*8*	7	6	2.1453	.00356
1	*2*	3	4	5	*8*	7	6	2.1453	.00356
2	1	3	4	5	*8*	6	7	2.8049	.00082
1	*2*	3	4	5	*8*	6	7	2.8050	.00082
8	*1*	2	3	4	5	6	7	3.2372	.00024
2	1	3	4	5	*7*	6	8	4.0877	.00004
1	2	3	4	5	*7*	6	8	4.0881	.00004
8	2	1	3	4	5	6	7	4.1684	.00002
2	1	3	4	5	*7*	*8*	6	4.2325	.00003
1	*2*	3	4	5	*7*	*8*	6	4.2325	.00003
3	2	1	4	5	6	7	8	4.2826	.00002
3	*1*	2	4	5	6	7	8	4.2833	.00002
7	*8*	*1*	2	3	4	5	6	4.2873	.00002
7	*8*	2	1	3	4	5	6	4.4594	.00001
2	1	3	*5*	*4*	6	7	8	4.4606	.00002
1	*2*	3	*5*	*4*	6	7	8	4.4609	.00002
2	*3*	*1*	4	5	6	7	8	4.5038	.00002
1	*3*	2	4	5	6	7	8	4.5048	.00002

[a] Italics indicate rearrangements relative to the most likely order.

[b] $\Delta\log_{10}L$ indicates the difference in maximum log likelihood from the best order.

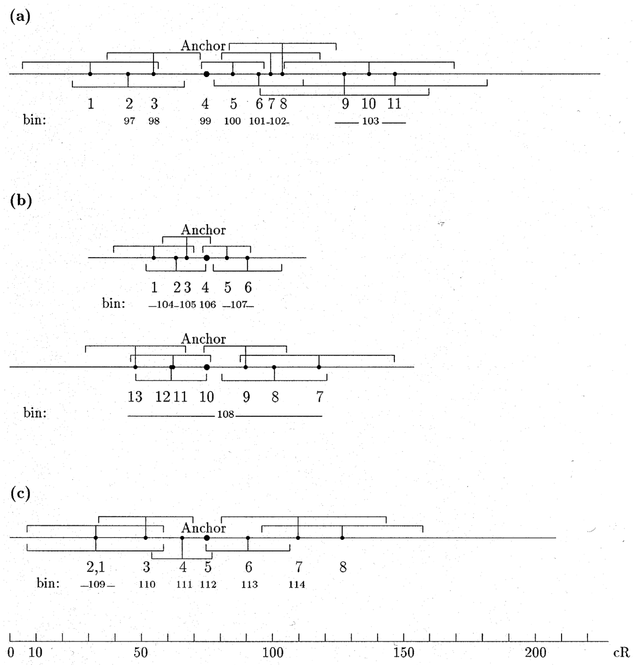

Figure 4 Maps of 32 markers on chromosome 1q. *a,* Map of chromosome 1q markers in linkage group 1. Prior map length $1.5 * \frac{12}{10}\hat{D} = 225$ cR, anchor = 4. *b,* Maps of chromosome 1q markers in linkage group 2: linkage group 2a—prior map length $1.5 * \frac{7}{5}\hat{D} = 83$ cR, anchor = 4; linkage group 2b—prior map length $1.5 * \frac{8}{6}\hat{D} = 154$ cR, anchor = 10. *c,* Map of chromosome 1q markers in linkage group 3. Prior map length $1.5 * \frac{9}{7}\hat{D} = 208$ cR, anchor = 5.

order (difference from best order $\Delta\log_{10}L = .0004$) flips markers 25 and 26 (fig. 4*c,* markers 1 and 2, bin 109). Our method illustrates this by placing the point estimates for these markers in nearly identical locations. (In contrast, the best maximum likelihood map places these markers >10 cR apart.) The third most likely order ($\Delta\log_{10}L = .0005$) flips markers 22 and 23 (fig. 4*b,* markers 11 and 12). Once again, this is illustrated in our method by point estimates that nearly coincide.

Although the widths of the posterior credible intervals may not be helpful in the ordering of markers, they can assist in determining which of a set of markers may be of interest for further studies. For example, the intervals can show which markers may lie between two flanking markers from a linkage study, or they may show which markers overlap, in location, with a particular marker of interest.

Our method should be most useful when there is no clearly best order for a set of markers, since it provides a way to order markers within a bin or within a framework map interval. However, even in situations in which there is a clearly best order, our method provides a better estimate of map length than maximum likelihood does. In either case, the interval estimates should be helpful in determining an appropriate set of markers for further study.

Acknowledgments

Support for this work was provided by research grants HG00376 (to M.B.) and GM53275 (to K.L.) and training grant T32 HG00040 (to H.M.S.) from the National Institutes of Health and by a University of Michigan Regents' Pre-Doctoral Fellowship (to H.M.S.). We thank David Cox and Kate McKusick for sharing the chromosome 1 data and a draft of their paper describing the SAMapper program.

Electronic-Database Information

The URL for data in this article is as follows:

Stanford Human Genome Center, http://www-shgc.stanford.edu/Mapping/rh/MapsV2/search1.html (for chromosome 1q marker bins)

References

Boehnke M, Lange K, Cox DR (1991) Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet 49: 1174–1188

Cox DR, Burmeister M, Price ER, Kim S, Myers RM (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science 250:245–250

de Bruijn NG (1981) Asymptotic methods in analysis. Dover Books, New York, pp 71–72

Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. Nature 255:680–684

Lange K (1997) Mathematical and statistical methods for genetic analysis. Springer-Verlag, New York, pp 35–36, 191–199

Lange K, Boehnke M, Cox DR, Lunetta KL (1995) Statistical methods for polyploid radiation hybrid mapping. Genome Res 5:136–150

Matise TC, Chakravarti A (1995) Automated construction of radiation hybrid maps using MultiMap. Am J Hum Genet Suppl 57:A15

Slonim D, Kruglyak L, Stein L, Lander E (1997) Building human genome maps with radiation hybrids. J Comput Biol 4:487–504

Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. J Am Stat Assoc 81: 82–86

Walter MA, Spillett DJ, Thomas P, Weissenbach J, Goodfellow PN (1994) A method for constructing radiation hybrid maps of whole genomes. Nat Genet 7:22–28

Xia T (1997) Ordering radiation hybrid markers using genetic algorithm. Am J Hum Genet Suppl 61:A247