Statistical methods for polyploid radiation hybrid mapping.

K Lange, M Boehnke, D R Cox, et al.

Genome Res. 1995 5: 136-150
Access the most recent version at doi:10.1101/gr.5.2.136

RESEARCH

Statistical Methods for Polyploid Radiation Hybrid Mapping

Kenneth Lange,[1,5] Michael Boehnke,[2] David R. Cox,[3] and Kathryn L. Lunetta[4]

[1,2,4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109; [3]Department of Genetics, School of Medicine, Stanford University, Stanford, California 94305

Radiation hybrid mapping is a somatic cell technique for ordering genetic loci along a chromosome and estimating physical distances between adjacent loci. This paper presents a model of fragment generation and retention for data involving two or more copies of the chromosome of interest per clone. Such polyploid data can be generated by initially irradiating normal diploid cells or by pooling haploid or diploid clones. The current model assumes that fragments are generated in the ancestral cell of a clone according to an independent Poisson breakage process along each chromosome. Once generated, fragments are independently retained in the clone with a common retention probability. On the basis of this and less restrictive retention models, statistical criteria such as minimum obligate breaks, maximum likelihood ratios, and Bayesian posterior probabilities can be used to decide locus order. Distances can be estimated by maximum likelihood. Likelihood computation is particularly challenging, and computing techniques from the theory of hidden Markov chains prove crucial. Within this context it is possible to incorporate typing errors. The statistical tools discussed here are applied to 14 loci on the short arm of human chromosome 4.

Radiation hybrids have proved to be a powerful and convenient method for rapidly mapping marker loci (Cox et al. 1990; Goss and Harris 1975). To date, most radiation hybrid maps have been constructed using haploid hybrids, which are generated by irradiating rodent cells carrying a single copy of a unique human chromosome. If the irradiated cells are fused with nonirradiated rodent cells, standard cell culture techniques permit selection for and eventual isolation of independent clones of hybrid cells. Each hybrid clone retains multiple random fragments from the unique human chromosome. One limitation of this experimental paradigm is that different hybrid panels must be prepared for each chromosome. In contrast, if human diploid cells are irradiated to generate diploid, "whole-genome" radiation hybrids as originally proposed by Goss and Harris (1977), then a single panel of hybrids can be used to map all human chromosomes.

In spite of the fact that multiple copies of a chromosome per clone obscures fragment retention patterns, diploid and polyploid radiation hybrids provide other advantages over haploid hybrids besides ease of generation. For instance, the mapping of closely spaced loci requires fragments of small average size. Such fragments may have low retention rates in cells. Using diploid clones or pooling haploid or diploid clones increases the effective retention rate per clone.

Any strategy for ordering loci from radiation hybrid data is necessarily complex (Boehnke et al. 1991; Lange and Boehnke 1992). Our philosophy is to look at radiation hybrid data from a variety of perspectives. A fundamental barrier is the shear number of orders that must be considered. For m loci, this number is either $m!/2$ or $m!$, depending on the symmetry of the retention model employed. Methods such as minimum obligate breaks involve simple criteria that allow rapid screening of many orders. Other methods such as maximum likelihood and Bayesian posterior probabilities are more computationally intensive and involve more modeling assumptions but provide a more satisfactory basis for comparing locus orders and estimating distances between loci.

At first glance, polyploid data appear to be much more difficult to analyze than haploid data. Fortunately, this is not the case if one adopts the computational framework of hidden Markov chains (Baum 1972; Devijver 1985; Rabiner 1989) for maximum likelihood estimation.

[5]Corresponding author.
E-MAIL klange@umich.edu; FAX (313) 763-2215.

5:136–150 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

Even within this context, Bayesian calculations still seem intimidating. One remedy is to apply Laplace's approximation (de Bruijn 1981; Tierney and Kadane 1986; Barndorff-Nielsen and Cox 1989). This transforms the problem of computing posterior probabilities by numerical integration to one of finding the posterior mode by optimization. The necessary ideas behind these computational advances are developed fully in this paper. The corresponding software package, RHMAP version 2.01, is available free of charge from Michael Boehnke.

Models for Polyploid Radiation Hybrid Mapping

The rationale behind radiation hybrid mapping is simple. The closer two loci are together on a human chromosome, the less likely it is that radiation will cause a break between them. Thus, close loci will tend to be concordantly retained or lost, whereas distant loci will tend to be independently retained or lost. To flesh out this intuitive insight, we make six reasonable modeling assumptions.

First, the loci to be mapped are linearly arranged along a given human chromosome, which we identify with a line segment. Second, each clone contains fragments derived from c copies of this chromosome. The values $c = 1$ and $c = 2$ correspond to haploid and diploid hybrids, respectively. The term polyploid hybrid covers an arbitrary number of chromosome copies $c \geq 2$ per clone. Unless aneuploid cells are irradiated, clones must be pooled to attain a value of $c > 2$. For the sake of brevity, we will always refer to the sampling unit in a radiation hybrid experiment as a clone, whether it corresponds to a single clone or a pool of clones. Third, we assume that the breaks caused by radiation along any chromosome occur according to a Poisson process. These Poisson breakage processes are independent from chromosome to chromosome and identically distributed on homologous chromosomes. Fourth, fragments within a clone are retained and lost independently. As noted below, different fragments can be retained with different rates, but the retention processes are again independent and identically distributed from chromosome to chromosome. Fifth, breakage and retention operate independently of each other. Sixth, only the presence and not the number of markers in a clone can be detected at any locus.

These assumptions can be manipulated mathematically upon adopting appropriate notation. Let the number of loci to be ordered be m and the number of hybrid clones tested at these loci be h. Observations on the clones can be arranged in an $h \times m$ matrix $X = (X_{jk})$ such that X_{jk} takes the value 0 or 1, according as no markers or one or more markers are observed at the kth locus of the jth clone. If the typing results at this locus are missing or ambiguous, we set $X_{jk} = ?$. The physical distance between loci k and $k + 1$ is denoted δ_k. If we assume that the independent Poisson breakage process are homogeneous with common intensity λ, then $d_k = \lambda\delta_k$ is the expected number of breaks between the two loci per chromosome copy. The probability of at least one break occurring between the loci on a given chromosome is

$$\theta_k = 1 - e^{-d_k}. \quad (1)$$

The breakage probabilities θ_k or the scaled distances d_k are more convenient parameters than the physical distances δ_k. For one thing, the intensity λ and the physical distances δ_k are confounded; for another, using the θ_k or d_k as parameters relieves us of making the assumption that the identically distributed Poisson breakage processes are homogeneous.

One can pose a number of retention models incorporating independent retention of fragments. The simplest postulates a common retention probability r for all fragments. Because there is abundant evidence (Cox et al. 1990; Ceccherini et al. 1992, Gorski et al. 1992) indicating that fragments bearing a centromere are preferentially retained, it is helpful to elaborate this model slightly (Boehnke et al. 1991). Suppose that the m loci are arranged in numerical order from left to right along a single chromosome arm with locus 1 closest to the centromere. It seems reasonable to postulate two distinct retention probabilities, one for those fragments containing locus 1 and one for those fragments not containing locus 1. Slightly more general than this centromeric model is the model assuming a distinct retention probability r_k for the class of fragments beginning to the left of locus k and to the right of locus k–1. This is the most general model consistent with rapid calculation of likelihoods. Unless stated otherwise, we will employ this left-endpoint model with m retention probabilities $r_1, \ldots, r_m$.

Pairwise Criteria for Deciding Order

Computing the number of obligate breaks per order allows comparisons of different orders

(Boehnke 1992; Boehnke et al. 1991; Bishop and Crockford 1992; Weeks et al. 1992). To illustrate the basic idea, the clone (0,0,0,0,0,1,0,0,0,0,?,0,0,1) from the chromosome 21 haploid data of Cox et al. (1990) displays three obligate breaks. Here we assume that the order of the loci along the chromosome is the same as the scoring order. Obligate breaks occur whenever a run of 0's is ended by a 1 or vice versa; untyped loci are ignored in this accounting. If the number of obligate breaks per clone is summed over all clones, then the resulting sum serves as a criterion for comparing the current order to other orders. Among the $m!/2$ possible orders, a best order or orders can be identified by minimizing the obligate breaks criterion using a stepwise ordering algorithm (Boehnke et al. 1991) or standard combinatorial optimization techniques such as branch-and-bound (Nijenhuis and Wilf 1978) and simulated annealing (Press et al. 1992).

The advantage of the minimum breaks criterion is that it depends on almost no assumptions about how breaks occur and fragments are retained. Under our specific model assumptions, the criterion is also statistically consistent given a common retention probability. Building on previous work of Barrett (1992) and Speed et al. (1992), we prove this claim in the Appendix. With minor notational changes, the proof in the Appendix shows that minimizing the estimated total map length or the estimated sum of adjacent breakage probabilities between the first and last loci of an order σ also provides strongly consistent criteria for choosing the true order. If $\hat{\theta}^h_{jk}$ is any strongly consistent sequence of estimators of the breakage probability θ_{jk} between two loci j and k, then these criteria can be expressed as

$$-\sum_{k=1}^{m-1} \ln[1 - \hat{\theta}^h_{\sigma(k),\sigma(k+1)}] \tag{2}$$

and

$$\sum_{k=1}^{m-1} \hat{\theta}^h_{\sigma(k),\sigma(k+1),} \tag{3}$$

respectively. The two-locus maximum likelihood estimates proposed in the next section are strongly consistent. Because the total map length (equation 2) and the sum of adjacent breakage probabilities (equation 3) are additive functions requiring one term for each pair of adjacent loci, a best order can be identified by the same techniques employed in minimizing obligate breaks. We tend to prefer the minimum of equation 3 to the minimum of equation 2 as a criterion for deciding order because equation 3 is less sensitive to errors in estimating large breakage probabilities (Olson and Boehnke 1990).

Likelihoods for One- and Two-locus Models

It is instructive to begin our discussion of maximum likelihood methods by considering one- and two-locus data from a single, fully typed clone. To simplify notation, we drop row subscripts and let $X = (X_1, \ldots, X_m)$ denote the observation vector for the clone. Recall that if no markers are present at locus k, then $X_k = 0$. If one or more markers are present, then $X_k = 1$. For a single locus, the only parameter of interest is the retention probability r, from which the subscript can also be dropped. Because $(1 - r)^c$ is the probability that all c copies of a given marker are lost, the single-locus polyploid likelihood reduces in the absence of typing errors to the Bernoulli distribution

$$q_0 = \Pr(X_1 = 0) = (1 - r)^c$$
$$q_1 = \Pr(X_1 = 1) = 1 - (1 - r)^c.$$

For two loci we again assume complete data and a common retention probability r. With the abbreviations $\theta = \theta_1$ and $q_{kl} = \Pr(X_1 = k, X_2 = l)$, two-locus polyploid likelihoods can be written as

$$\begin{aligned} q_{00} &= [(1 - r)(1 - \theta r)]^c \\ q_{10} &= q_{01} \\ &= q_0 - q_{00} \\ &= (1 - r)^c - [(1 - r)(1 - \theta r)]^c \\ q_{11} &= 1 - q_{00} - q_{10} - q_{01} \\ &= 1 - 2(1 - r)^c + [(1 - r)(1 - \theta r)]^c. \end{aligned} \tag{4}$$

The expression for $q_{00} = \Pr(X_1 = 0, X_2 = 0)$ in equation 4 is a direct consequence of the independent fate of the c chromosomes during fragmentation and retention. Considering a given chromosome, the marker at locus 1 is lost with probability $1 - r$. Conditional on this event, the marker at locus 2 is lost with probability $1 - \theta r$ because the complementary event occurs only when there is a break between the two loci and the fragment bearing the second locus is retained.

The parameters r and θ can be expressed in terms of the two probabilities q_{00} and q_{11}. The

equation $q_{11} - q_{00} = 1 - 2(1 - r)^c$ can be solved to give

$$r = 1 - \left[\frac{1 - q_{11} + q_{00}}{2}\right]^{\frac{1}{c}}. \quad (5)$$

When $c = 1$, this simplifies to $r = q_{11} + q_{10}$. Once r is determined, solving for θ in $q_{00} = [(1 - r)(1 - \theta r)]^c$ gives

$$\theta = \frac{1 - r - [q_{00}]^{\frac{1}{c}}}{r(1 - r)}. \quad (6)$$

Again when $c = 1$, this reduces to $\theta = (q_{10})/[r(1 - r)]$. In general the two-dimensional map $(\theta,r) \rightarrow (q_{11},q_{00})$ is one-to-one from the region

$$\{(\theta,r):\theta \in [0,1],\ r \in (0,1)\}$$

onto the region

$$R = \{(q_{00},q_{11}):q_{00} \in (0,1),\ q_{11} \in (0,1),\ q_{00}q_{11} \geq q_{10}^2\}.$$

Furthermore, $\theta = 0$ if and only if $q_{00} + q_{11} = 1$, and $\theta = 1$ if and only if $q_{00}q_{11} = q_{10}^2$. The upper boundary of the region R is formed by the line $q_{00} + q_{11} = 1$ and the lower boundary by the curve $q_{00}q_{11} = q_{10}^2$, which in turn is generated by the function $q_{11} = 1 + q_{00} - 2\sqrt{q_{00}}$.

The observed values of q_{11} and q_{00} are maximum likelihood estimates for the simplified multinomial model in which the only constraints on the four probabilities q_{00}, q_{10}, q_{01}, and q_{11} are non-negativity, the symmetry condition $q_{10} = q_{01}$, and the sum requirement $q_{00} + q_{10} + q_{01} + q_{11} = 1$. This simplified model has in effect two parameters, which we can identify with q_{11} and q_{00} and estimate by their empirical values. These values are maximum likelihood estimates under the simplified model. If these estimates satisfy the inequality $q_{00}q_{11} \geq q_{10}^2$, then they furnish maximum likelihood estimates for the radiation hybrid model as well. Because maximum likelihood estimates are preserved under reparameterization, the maximum likelihood estimates of r and θ are then available by substituting estimated values for theoretical values in equations 5 and 6. In the event that the estimated q's do not satisfy $q_{00}q_{11} > q_{10}^2$, then it is prudent to set θ equal to some number slightly less than 1. For more than two loci, the two-locus maximum likelihood estimates furnish good starting values for the breakage probabilities in a full maximum likelihood analysis.

The above analysis depends crucially on the assumption of a common retention probability r at the two loci. A reasonable diagnostic for this assumption is to test the hypothesis $q_{10} = q_{01}$. The simplest approach is to use the clones discordant at the two loci and test whether the binomial distribution with success probability ½ fits the corresponding observed numbers. A poor fit would favor application of the left-endpoint model for fragment retention. In any case, such exploratory data snooping is a good preliminary to more complex modeling.

Typing Errors

One method of controlling typing errors is to retest every clone at every locus. If the two tests agree, then the clone is assigned the common result at the locus. If the tests differ, then the clone is viewed as untyped at the locus. This procedure ensures a very low error rate. Instead of retyping, the less expensive, but less reliable option of rescoring is available. As an alternative to retyping and rescoring, one can construct a model that specifically incorporates the possibility of typing error. Both false-negative and false-positive errors should be taken into account.

For loci typed by the polymerase chain reaction, false negatives arise from a failure of target DNA to amplify sufficiently. False positives arise from nonspecific DNA amplification, which we will refer to as "contamination." Let α_k be the amplification rate and β_k be the contamination rate at locus k. Perfect amplification corresponds to $\alpha_k = 1$ and no contamination to $\beta_k = 0$. Borrowing a term from pedigree analysis, we incorporate these parameters into the penetrance expression for the clone at locus k (Lazzeroni et al. 1994). If G_k denotes the random number of markers present in the clone at this locus, then the penetrance $\phi_k(x_k \mid g_k)$ is the conditional probability of the observation $X_k = x_k$ given $G_k = g_k$. It is simplest to assume that typing results are independent from locus to locus given the $G_1, \ldots, G_m$; in other words,

$$\Pr(X_1 = x_1, \ldots, X_m = x_m \mid G_1 = g_1, \ldots, G_m = g_m) = \prod_{k=1}^{m} \phi_k(x_k \mid g_k).$$

It is also reasonable to assume independence of amplification and contamination events within a locus. This assumption makes it feasible to consider more complicated typing schemes. For instance, multitest penetrances can be formed by

multiplying the underlying single-test penetrances.

Two amplification models are plausible. If each marker copy at a locus amplifies independently, then

$$\phi_k(0 \mid g_k) = (1 - \alpha_k)^{g_k}(1 - \beta_k). \tag{7}$$

In contrast, if amplification is an all-or-none phenomenon, then

$$\phi_k(0 \mid g_k) = (1 - \alpha_k)^{\min\{g_k,1\}}(1 - \beta_k). \tag{8}$$

Of course, $\phi_k(1 \mid g_k)$ is determined by the relation $\phi_k(0 \mid g_k) + \phi_k(1 \mid g_k) = 1$. A necessary convention for missing data is $\phi_k(? \mid g_k) = 1$.

Alternatively, one can model typing errors in a slightly less mechanistic way by simply postulating a false positive rate $\beta_k = \phi_k(1 \mid 0)$ and a false-negative rate $\nu_k = \phi_k(0 \mid g_k)$ for $g_k \geq 1$. This is equivalent to the all-or-none amplification model with $\nu_k = (1 - \alpha_k)(1 - \beta_k)$ except that the implicit constraint $\nu_k \leq 1 - \beta_k$ no longer applies. The false-positive and false-negative parameterization is slightly less attractive for independent amplification as exemplified in equation 7.

In practice, the parameters α_k and β_k (or ν_k and β_k) can be fixed at reasonable values or estimated from the data. If they are estimated from the data, there are two obvious possibilities. As implied by our notation, α_k and β_k can be locus specific. Another possibility is to estimate a single amplification rate $\alpha_k = \alpha$ and a single contamination rate $\beta_k = \beta$ common to all loci. In principle, making error rates locus specific permits identification of loci with poor typing results. One may want to rescore, retype, or simply discard such loci before forming a map.

A Hidden Markov Chain for Likelihood Calculation

Maximum likelihood estimation is considerably more difficult for multiple loci than for one or two loci. Nonetheless, it is possible to design a fast algorithm for likelihood calculation based on the theory of hidden Markov chains (Baum 1972; Devijver 1985; Rabiner 1989). This new algorithm also handles missing data and typing errors gracefully. To derive the algorithm, we posit a Markov chain for the current clone whose state G_k at locus k is the number of copies of marker k present in the clone. As the chain progresses from locus k to locus $k + 1$, starting at locus 1 and ending at locus m, G_k is updated to G_{k+1}. The numbers G_k are hidden from view because only the presence or absence of markers are directly observable. Of fundamental importance in understanding the chain are the probabilities $\Pr(G_{k+1} = j \mid G_k = i) = t_{c,k}(i,j)$ of a transition from state i at locus k to state j at locus $k + 1$.

To compute $t_{c,k}(i,j)$, consider first a haploid clone. In this situation the chromosome copy number $c = 1$, and it is clear from our earlier analysis that

$$\begin{aligned} t_{1,k}(0,0) &= 1 - \theta_k r_{k+1} \\ t_{1,k}(0,1) &= \theta_k r_{k+1} \\ t_{1,k}(1,0) &= \theta_k(1 - r_{k+1}) \\ t_{1,k}(1,1) &= 1 - \theta_k(1 - r_{k+1}). \end{aligned}$$

Employing these haploid transition probabilities, we can write the following general expression

$$\begin{aligned} t_{c,k}(i,j) = \sum_{l=\max\{0,i+j-c\}}^{\min\{i,j\}} &\binom{i}{l} t_{1,k}(1,1)^l t_{1,k}(1,0)^{i-l} \\ &\times \binom{c-i}{j-l} t_{1,k}(0,1)^{j-l} t_{1,k}(0,0)^{c-i-j+l} \end{aligned} \tag{9}$$

for the polyploid transition probabilities. Formula 9 can be deduced by letting l be the number of markers retained at locus k that lead via the same original chromosomes to markers retained at locus $k + 1$. These l markers can be chosen in $\binom{i}{l}$ ways. Among the i markers retained at locus k, the fate of the l markers retained at locus $k + 1$ and the remaining $i - l$ markers not retained at locus $k + 1$ is captured by the product $t_{1,k}(1,1)^l t_{1,k}(1,0)^{i-l}$. For j total markers to be retained at locus $k + 1$, the $c - i$ markers not retained at locus k must lead to $j - l$ markers retained at locus $k + 1$. These $j - l$ markers can be chosen in $\binom{c-l}{j-l}$ ways. The product $t_{1,k}(0,1)^{j-l} t_{1,k}(0,0)^{c-i-j+l}$ captures the fate of the $c - i$ markers not retained at locus k. Finally, the upper and lower bounds on the index of summation l ensure that none of the powers of the $t_{1,k}(u,v)$ appearing in 9 are negative.

The likelihood of the observations $(X_1, \ldots, X_m)$ from a clone can be written as

$$\begin{aligned} P &= \Pr(X_1 = x_1, \ldots, X_m = x_m) \\ &= \sum_{g_1} \cdots \sum_{g_m} \binom{c}{g_1} r_1^{g_1}(1 - r_1)^{c-g_1} \\ &\quad \prod_{k=1}^{m-1} t_{c,k}(g_k, g_{k+1}) \prod_{k=1}^{m} \phi_k(x_k \mid g_k). \end{aligned} \tag{10}$$

Expression 10 reflects the assumption that the observations $(X_1, \ldots, X_m)$ are independent given the underlying marker counts $(G_1, \ldots, G_m)$. In the absence of typing errors, the range of summation for the index g_k can be reduced to $g_k \in \{0\}$ for $x_k = 0$ and to $g_k \in \{1, \ldots, c\}$ for $x_k = 1$. For a single chromosome, this simplification implies that the likelihood factors (Boehnke et al. 1991). In the polyploid case, this advantage disappears, but fast evaluation of the multiple sum (expression 10) as an iterated sum is still possible based on Baum's algorithms from the theory of hidden Markov chains (Baum 1972; Devijver 1985; Rabiner 1989).

Baum's forward algorithm recursively evaluates the joint probabilities

$$f_k(g_k) = \Pr(X_1 = x_1, \ldots, X_{k-1} = x_{k-1}, G_k = g_k)$$

beginning with the initial condition

$$f_1(g_1) = \Pr(G_1 = g_1) = \binom{c}{g_1} r^{g_1}(1 - r_1)^{c-g_1}$$

at the first locus. The forward update is

$$f_{k+1}(g_{k+1}) = \sum_{g_k} f_k(g_k)\phi_k(x_k \mid g_k)t_{c,k}(g_k, g_{k+1}).$$

The likelihood (expression 10) can be recovered by forming the sum

$$P = \sum_{g_m} f_m(g_m)\phi_m(x_m \mid g_m)$$

at the last locus. In the absence of typing errors, the various sums defining the forward algorithm range only over those marker counts consistent with the observed data.

Baum's backward algorithm recursively evaluates the conditional probabilities

$$b_k(g_k) = \Pr(X_{k+1} = x_{k+1}, \ldots, X_m = x_m \mid G_k = g_k)$$

starting by convention at $b_m(g_m) = 1$. The required update is

$$b_{k-1}(g_{k-1}) = \sum_{g_k} t_{c,k-1}(g_{k-1}, g_k)\phi_k(x_k \mid g_k)b_k(g_k).$$

In this instance the likelihood can be recovered at the first locus by forming the sum $P = \Sigma_{g_1} f_1(g_1)\phi(x_1 \mid g_1)b_1(g_1)$.

A quick search of the likelihood P can be achieved via the EM algorithm (Dempster et al. 1977) if the partial derivatives of the likelihood can be computed analytically. Let us now indicate briefly how to do this based on the intermediate results of Baum's forward and backward algorithms. Consider first the partial derivative

$$\frac{\partial}{\partial \theta_k} P$$

of P with respect to a breakage probability θ_k. To isolate θ_k, we rewrite the likelihood (expression 10) as

$$P = \sum_{g_k} \sum_{g_{k+1}} f_k(g_k)\phi_k(x_k \mid g_k)t_{c,k}(g_k, g_{k+1}) \times \phi_{k+1}(x_{k+1} \mid g_{k+1})b_{k+1}(g_{k+1}). \quad (11)$$

Differentiating expression 11 with respect to θ_k yields

$$\frac{\partial}{\partial \theta_k} P = \sum_{g_k} \sum_{g_{k+1}} f_k(g_k)\phi_k(x_k \mid g_k) \frac{\partial}{\partial \theta_k} t_{c,k}(g_k, g_{k+1}) \times \phi_{k+1}(x_{k+1} \mid g_{k+1})b_{k+1}(g_{k+1}). \quad (12)$$

An analogous formula is valid for each r_k except that

$$\frac{\partial}{\partial r_1} P = \sum_{g_1} \frac{\partial}{\partial r_1} f_1(g_1)\phi_1(x_1 \mid g_1)b_1(g_1). \quad (13)$$

For an amplification or contamination parameter $\gamma_k = \alpha_k$ or β_k,

$$\frac{\partial}{\partial \gamma_k} P = \sum_{g_k} f_k(g_k) \frac{\partial}{\partial \gamma_k} \phi_k(x_k \mid g_k)b_k(g_k). \quad (14)$$

If several parameters are consolidated into a single parameter, then the chain rule must be applied. For instance, if a single retention probability r is assumed, then

$$\frac{\partial}{\partial r} P = \sum_k \frac{\partial}{\partial r_k} P.$$

The partial derivatives appearing on the right hand sides of expressions 12, 13, and 14 are tedious, but straightforward to evaluate. Efficient evaluation of P and its partial derivatives can be orchestrated by carrying out the backward algorithm first, followed by the forward algorithm performed simultaneously with the computation of all partial derivatives. Given a partial derivative

$$\frac{\partial}{\partial \gamma_k} P$$

of the likelihood P, one forms the corresponding entry

$$\frac{\partial}{\partial \gamma_k} \ln P$$

in the score vector by taking the quotient

$$\frac{\frac{\partial}{\partial\gamma_k}P}{P}.$$

Likelihood Maximization

The EM algorithm provides an attractive avenue to maximum likelihood estimation of the model parameters (Dempster et al. 1977). Suppose that we collect the parameters into a vector γ with typical entry γ_k. Each of the γ_k can be viewed as a success probability for a hidden binomial trial. As a consequence (Weeks and Lange 1989), the EM update for any parameter takes either of the equivalent generic forms

$$\begin{aligned}\gamma_k^{n+1} &= \frac{E(\#success \mid X,\gamma^n)}{E(\#trials \mid X,\gamma^n)} \\ &= \gamma_k^n + \frac{\gamma_k^n(1-\gamma_k^n)\frac{\partial L(\gamma^n)}{\partial\gamma_k}}{E(\#trials \mid X,\gamma^n)}\end{aligned} \tag{15}$$

where X denotes the observations over all clones, and $L = \ln P$ is the log-likelihood function. The second form of the update requires less thought to implement, as we already know how to compute the partial derivatives $(\partial)/(\partial\gamma_k)L(\gamma^n)$. Because the EM algorithm will not budge from either $\gamma_k^n = 0$ or $\gamma_k^n = 1$, these boundary values should be avoided as initial points.

If there are h clones and γ_k is a breakage probability θ_k, then the conditional expectation $E(\#trials \mid X,\gamma^n)$ appearing in expression 15 equals the constant hc. For a contamination rate β_k, we have $E(\#trials \mid X,\gamma^n) = h$. For a retention probability r_k, this conditional expectation is more subtle to calculate. When $k > 1$, the number of trials coincides with the random number of breaks W_k between loci $k-1$ and k among all clones. The first form of the EM update in expression 15 for θ_{k-1} shows that

$$\theta_{k-1}^{n+1} = \frac{E(W_k \mid X,\gamma^n)}{hc}.$$

It follows that the expected number of fragments with retention probability r_k is $hc\theta_{k-1}^{n+1}$. The expected number of fragments for retention probability r_1 is again just the constant hc.

Calculation of $E(\#trials \mid X,\gamma^n)$ for an amplification rate α_k can best be achieved by direct computation. Under the independent amplification model, this conditional expectation for a single clone can be expressed as

$$\frac{\Sigma_{g_k} g_k f_k(g_k)\phi_k(x_k \mid g_k)b_k(g_k)}{P}. \tag{16}$$

Under the all-or-none amplification model, expression 16 should be modified by substituting $\min\{g_k,1\}$ for the number of marker copies g_k at locus k. If several parameters are consolidated into a single parameter, then $E(\#trials \mid X,\gamma^n)$ can be calculated for the consolidated parameter by summing the corresponding conditional expectations over the various contributing parameters.

Once the maximum likelihood estimate $\hat{\gamma}$ is computed under the best order, one can compute parameter asymptotic standard errors and correlations by inverting the observed information matrix, that is, the matrix of negative second partial derivatives of $L(\gamma)$ evaluated at $\gamma = \hat{\gamma}$. The observed information matrix can be computed by numerically differentiating the score vector. Based on maximum likelihood estimates under the best identified order, one can test by standard likelihood ratio methods hypotheses imposed on the parameters. For instance, it can be revealing to test whether amplification or contamination rates differ significantly from locus to locus.

Bayesian Methods

Because Bayesian methods directly yield posterior probabilities of locus order, they offer an attractive alternative to maximum likelihood methods. To implement a Bayesian analysis, two technical hurdles must be overcome. First, an appropriate prior distribution for the parameters must be chosen. Once this choice is made, efficient numerical schemes for estimating parameters and posterior probabilities must be constructed.

It is more convenient to put a prior on the distances between the adjacent loci of an order than on the breakage probabilities determined by these distances. To specify a prior, we assume that the m loci to be mapped are sampled uniformly from a chromosome interval of known physical length. One can exploit the outcome of a previous maximum likelihood analysis of the hybrid data to assign to the prior interval a distance measured in expected number of breaks per chromosome. (The units on this distance are rays or centirays (cR) = 100 × rays.) Suppose that under the best maximum likelihood order, we esti-

mate a total of b expected breaks between the first and last locus. With m uniformly distributed loci, adjacent pairs of loci should be separated by an average distance of $b/(m-1)$ expected breaks. This quantity should also approximate the average distance from the left end of the interval to the first locus and from the right end of the interval to the last locus. These considerations suggest that $d = (m+1)b/(m-1)$ would be a reasonable expected number of breaks to assign to the prior interval. In practice, this value of d may be too confining, and it is probably prudent to inflate it by 10%–20%.

Given a prior interval of length d and a given locus order, let d_k be the distance separating the adjacent loci k and $k+1$. If the loci are uniformly and independently distributed on the interval, then standard arguments from geometric probability (Feller 1971) imply that the vector $(d_1, \ldots, d_{m-1})$ has prior density

$$\frac{m!(d - d_1 - \cdots - d_{m-1})}{d^m} \tag{17}$$

on the set

$$\{(d_1,\ldots, d_{m-1}):0 \leq d_k,\ 1 \leq k \leq m-1,\ \sum_{k=1}^{m-1} d_k \leq d\}.$$

Independent beta priors

$$\frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \gamma_k^{a_k-1}(1-\gamma_k)^{b_k-1} \tag{18}$$

can be reasonably assigned to the remaining retention, amplification, and contamination parameters. The beta family is flexible enough to include many differently shaped densities on [0,1]. For instance, taking $a_k = b_k = 1$ gives a flat prior with mean ½. The general β density has mean $a_k/(a_k + b_k)$.

With the resulting product prior now fixed for the full parameter vector γ, we can estimate parameters by maximizing the log posterior $L(\gamma) + R(\gamma)$, where $L(\gamma)$ is the log-likelihood and $R(\gamma)$ is the sum of the logarithms of the distance prior (expression 17) and of the retention, amplification, and contamination priors (expression 18). This maximization provides the posterior mode $\hat{\gamma}$. Although the EM algorithm is again our preferred method of optimization, the M (maximization) step is now only partially tractable. In the E step of the classical EM algorithm, one forms the conditional expectation $Q(\gamma \mid \gamma^n)$ of the complete data loglikelihood with respect to the observed data; in the M step one then maximizes $Q(\gamma \mid \gamma^n)$ as a function of its left argument γ. In the Bayesian EM algorithm one maximizes the amended function $Q(\gamma \mid \gamma^n) + R(\gamma)$. This surrogate function for the log posterior separates the retention, amplification, and contamination parameters, but the distance parameters are inextricably tied together through the prior density (expression 17). By analogy to the classical EM updates (expression 15), the Bayesian EM updates for the retention, amplification, and contamination parameters all reduce to either of the two equivalent expressions:

$$\begin{aligned}\gamma_k^{n+1} &= \frac{E(\#successes \mid X,\gamma^n) + a_k - 1}{E(\#trials \mid X,\gamma^n) + a_k + b_k - 2} \\ &= \gamma_k^n + \frac{\gamma_k^n(1-\gamma_k^n)\left[\dfrac{\partial L(\gamma^n)}{\partial \gamma_k} + \dfrac{\partial R(\gamma^n)}{\partial \gamma_k}\right]}{E(\#trials \mid X,\gamma^n) + a_k + b_k - 2}\end{aligned} \tag{19}$$

for hidden binomial trials, provided $E(\#successes \mid X,\gamma^n) + a_k - 1$ and $E(\#trials \mid X,\gamma^n) + a_k + b_k - 2$ are both positive. These positivity constraints certainly hold under the reasonable assumptions that $a_k > 1$ and $b_k > 1$.

The M step for the distance parameters can be well approximated by a modification of the EM algorithm known as the EM gradient algorithm (Lange 1995). The full EM gradient algorithm updates γ via

$$\begin{aligned}\gamma^{n+1} = \gamma^n - [d^{20}Q(\gamma^n \mid \gamma^n) + d^2R(\gamma^n)]^{-1} \\ \times [dL(\gamma^n) + dR(\gamma^n)]^t,\end{aligned} \tag{20}$$

where dL and dR denote the differentials of L and R, d^2R is the second differential of R, $d^{20}Q(\gamma \mid \gamma^n)$ is the second differential of Q relative to its left argument, and the superscript t represents vector transpose. In essence, the EM gradient algorithm approximately maximizes $\gamma \to Q(\gamma \mid \gamma^n) + R(\gamma)$ by one step of Newton's method. Maximizing this surrogate function forces the desired increase $L(\gamma^{n+1}) + R(\gamma^{n+1}) > L(\gamma^n) + R(\gamma^n)$. Note that the identity $d^{10}Q(\gamma^n \mid \gamma^n) = dL(\gamma^n)$ is employed in writing the Newton update as expression 20.

In this particular problem, it is unnecessary to apply the EM gradient update (expression 20) to all of the parameters. The exact update (expression 19) is applied where possible, and the EM gradient update is reserved for the distance parameters. Thus, we interpret the first differentials dL and dR appearing in expression 20 as $(m-1) \times 1$ row vectors pertaining only to the distance parameters. The second differentials

$d^{20}Q$ and d^2R we likewise interpret as $(m-1)\times(m-1)$ matrices.

All of the terms appearing in expression 20 are straightforward to evaluate. For instance, taking into account relation 1, we have

$$\frac{\partial}{\partial d_k}L(\gamma)=\frac{\partial}{\partial\theta_k}L(\gamma)\frac{d\theta_k}{dd_k}$$
$$=\frac{\partial}{\partial\theta_k}L(\gamma)(1-\theta_k).$$

Differentiation of $R(\gamma)$ with respect to the interlocus distances yields

$$\frac{\partial}{\partial d_\kappa}R(\gamma)=-\frac{1}{d-d_1-\cdots-d_{m-1}}$$
$$\frac{\partial^2}{\partial d_j d_k}R(\gamma)=-\frac{1}{(d-d_1-\cdots-d_{m-1})^2}.$$

Clearly, the matrix $d^2R(\gamma)$ is rank one.

The $(m-1)\times(m-1)$ Hessian matrix $d^{20}Q(\gamma\mid\gamma)$ is diagonal. To calculate one of its typical diagonal terms, again consider the random number W_{k+1} of chromosomes in the sample with breaks between loci k and $k+1$. As noted earlier, this random variable has a binomial distribution with success probability θ_k and hc trials. Now execution of the E step of the EM algorithm shows that up to an irrelevant constant,

$$Q(\gamma\mid\gamma^n)$$
$$=E(W_{k+1}\mid X,\gamma^n)\ln(\theta_k)+E(hc-W_{k+1}\mid X,\gamma^n)\ln(1-\theta_k).$$

Repeated differentiation with respect to the left variable of $Q(\gamma\mid\gamma^n)$ and then application of the chain rule yield

$$\frac{\partial^2}{\partial d_k^2}Q(\gamma\mid\gamma^n)=-\frac{E(W_{k+1}\mid X,\gamma^n)(1-\theta_k)}{\theta_k^2}.$$

The value $E(W_{k+1}\mid X,\gamma^n)$ has already been specified in our discussion of the EM algorithm in the absence of a prior.

The fact that the matrix $-[d^{20}Q(\gamma^n\mid\gamma^n)+d^2R(\gamma^n)]$ is a rank-one perturbation of a diagonal matrix is helpful. For such matrices, matrix inversion is easy owing to the availability of the Sherman-Morrison formula

$$(A+uu^t)^{-1}v=A^{-1}v-\frac{u^tA^{-1}v}{1+u^tA^{-1}u}A^{-1}u$$

for an invertible matrix A and compatible vectors u and v (Miller 1987).

Perhaps more important from the Bayesian perspective than finding the posterior mode is the possibility of computing posterior probabilities for the various locus orders. Under the natural assumption that all orders are a priori equally likely, the posterior probability of a given order ω is

$$\frac{\int e^{L_\omega(\gamma)+R_\omega(\gamma)}d\gamma}{\Sigma_\nu\int e^{L_\nu(\gamma)+R_\nu(\gamma)}d\gamma},\qquad(21)$$

where the sum in the denominator ranges over all possible orders ν and L_ν and R_ν denote the loglikelihood and log prior appropriate to order ν. Two ugly issues rear their heads immediately at this point. First, unless the number of loci m is small, the number of possible orders $m!$ or $m!/2$ can be astronomical. This problem can be finessed if the leading orders can be identified, for example, by minimum obligate breaks, and the sum truncated to include only these orders. In many problems only a few orders contribute substantially to the denominator of the posterior probability (equation 21).

The other issue is how to evaluate the integrals appearing in equation 21. Owing to the complexity of the integrands, there is no obvious analytic method of carrying out the integrations. For haploid data, Lange and Boehnke (1992) suggest two approximate methods. Both of these methods have their drawbacks and can be computationally demanding. Here we suggest an approximation based on Laplace's method from asymptotic analysis (de Bruijn 1981; Tierney and Kadane 1986; Barndorff-Nielsen and Cox 1989). The idea is to expand the logarithm of the integrand $e^{L_\omega(\gamma)+R_\omega(\gamma)}$ in a second-order Taylor's series around the posterior mode $\hat\gamma$. Recalling the well-known normalized constant for the multivariate normal density and defining $F_\omega(\gamma)=L_\omega(\gamma)+R_\omega(\gamma)$, this approximation yields

$$\int e^{F_\omega(\gamma)}d\gamma\approx\int e^{F_\omega(\hat\gamma)+\frac{1}{2}(\gamma-\hat\gamma)^t d^2F_\omega(\hat\gamma)(\gamma-\hat\gamma)}d\gamma$$
$$=e^{F_\omega(\hat\gamma)}(2\pi)^{\frac{m}{2}}\det(-d^2F_\omega(\hat\gamma))^{-\frac{1}{2}}.\qquad(22)$$

The accuracy of Laplace's approximation increases as the log posterior function becomes more peaked around the posterior mode $\hat\gamma$. The quadratic form $d^2F_\omega(\hat\gamma)$ measures the curvature of $F_\omega(\gamma)$ at $\hat\gamma$.

Application to Chromosome 4 Data

To illustrate some of the techniques introduced above, we now consider data on 14 sequence-tagged sites from the short arm of human chromosome 4. These markers, which are listed in Table 1, constitute a small subset of a much larger

Table 1. Chromosome 4 sequence-tagged site markers

Marker	Name	Marker	Name
1	STS4-475	8	STS4-842
2	STS4-163	9	STS4-161
3	STS4-555	10	STS4-543
4	STS4-5	11	STS4-879
5	STS4-259	12	STS4-759
6	STS4-246	13	STS4-476
7	STS4-613	14	STS4-547

set of markers typed on 83 radiation hybrids constructed at the Stanford University Human Genome Center and distributed by Research Genetics (Cox, D.R., R.M. Myers, D. Vollrath, M. Boehnke, and K. Lange, in prep.). The duplicate typing of these whole-genome, diploid hybrids make them ideal for exploring the nature of typing errors. In analyzing the combined data, we allowed for typing errors by the simple device of treating discordant markers within a clone as untyped. When analyzing the results of the two separate typings—referred to as gels 1 and 2—we used the more elaborate models that explicitly take into account typing errors.

The paramount concerns in any analysis of radiation hybrids are to identify the correct locus order and to estimate distances between loci under the best order. Table 2 lists for the combined data the 10 best maximum likelihood orders; these were found by the stepwise ordering algorithm described in Boehnke et al. (1991). These data show clear but not overwhelming support for the first order. The next order is plausible, but after it there is a marked decline in support for the remaining orders. The maximum likelihood and minimum obligate breaks criteria rank the first three orders identically but disagree noticeably for subsequent orders. Figure 1 depicts map distances and maximum likelihood pairwise inversion ratios for the best order.

If we adopt the criterion of maximum posterior probability for ordering loci, then we need to be concerned whether we are approximating posterior probabilities well. The posterior probabilities presented in Table 3 were calculated under three approximations to the integrals $\int e^{L\omega(\gamma)+R\omega(\gamma)}d\gamma$ appearing in expression (21). The first approximation is

$$\int e^{L_\omega(\gamma)+R_\omega(\gamma)}\,d\gamma \propto e^{L_\omega(\hat{\gamma})+R_\omega(\hat{\gamma})} \qquad (23)$$

where $\hat{\gamma}$ is the maximum likelihood estimate, and where the logprior $R_\omega(\gamma)$ is taken as 0. The second approximation employs equation 23 with a log prior $R_\omega(\gamma)$ constructed from the interlocus distance prior (equation 17) and a flat prior on the common retention probability r. The posterior mode replaces the maximum likelihood estimate. The third approximation is just the second approximation modified by the Laplace correction factor as indicated in equation 22. For each of these approximations, the normalizing sum in the denominator of the posterior probability was truncated to include only those 17 orders whose maximum likelihoods were within a factor of 10^{-6} of the max imum likelihood of the best order.

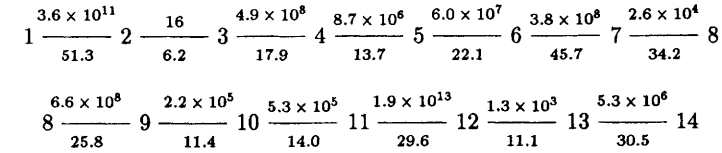

Figure 1 Map of the best order showing maximum likelihood ratios for pairwise inversions above and distance estimates in cR below.

Table 2. Best locus orders for the combined data

Rank	Orders	$\Delta\log_{10}L$[a]	Breaks
1	1-2-3-4-5-6-7-8-9-10-11-12-13-14	0.000	95
2[b]	1-**3-2**-4-5-6-7-8-9-10-11-12-13-14	1.218	97
3	1-2-3-4-5-6-7-8-9-10-11-**13-12**-14	3.120	99
4	2-3-4-5-6-7-8-9-10-11-12-13-14-**1**	3.355	108
5	**3-2**-4-5-6-7-8-9-10-11-12-13-14-1	3.795	108
6	7-8-9-10-11-12-13-14-**6-5-4-3-2-1**	4.216	106
7	1-**3-2**-4-5-6-7-8-9-10-11-**13-12**-14	4.339	101
8	1-2-3-4-5-6-**8-7**-9-10-11-12-13-14	4.408	101
9	**6-5-4-3-2-1**-7-8-9-10-11-12-13-14	5.076	102
10	1-2-3-4-5-6-7-8-9-10-11-**14-13-12**	5.147	103

[a] $\Delta\log_{10}L$: The difference in maximum log-likelihoods between the current order and order 1.
[b] Rearrangements relative to order 1 are noted in boldface for orders 2–10.

Table 3. Posterior probabilities for the combined data

ML rank[a] of order	Posterior probabilities		
	Approx. 1	Approx. 2	Approx. 3
1	0.94158	0.94795	0.94597
2	0.05679	0.05116	0.05312
3	0.00071	0.00062	0.00068
4	0.00042	0.00011	0.00008
5	0.00015	0.00004	0.00003
6	0.00006	0.00003	0.00003
7	0.00004	0.00003	0.00004
8	0.00004	0.00003	0.00003
9	0.00001	0.00000	0.00000
10	0.00001	0.00000	0.00001

[a]Orders ranked by maximum likelihood.

The good agreement displayed in Table 3 between the three methods of computing posterior probabilities seems to justify use of the simple maximum likelihood approximation, and all posterior probabilities quoted below involve this approximation.

Table 4 lists posterior probabilities for the same 10 orders based on analysis of gel 1 under six representative error models. (The results for gel 2 are similar.) These six models, labeled A–F, all invoke the all-or-none amplification mechanism indicated in equation 8. [Results for gel 1 under the independent amplification mechanism (equation 7) are very similar.] Model A assumes a single amplification rate and a single contamination rate with a flat prior on each. Model B assumes locus-specific amplification and contamination rates under flat priors. Models C and D duplicate models A and B, respectively, except that nonflat priors are assumed. For amplification rates, models C and D assume β parameters of $a = 19$ and $b = 1$, giving prior means of 0.95. For contamination rates, models C and D assume β parameters of $a = 1$ and $b = 99$, giving prior means of 0.01. Model E fixes all amplification rates at 0.995 and all contamination rates at 0.005. Finally, model F ignores typing errors and provides posterior probabilities comparable to those displayed in Table 3 for the combined data.

From the results in Table 4, it is clear that incorporating an explicit error model can reduce the posterior odds for the best order when this order is strongly supported by the data (Lunetta et al. 1995). Taking into account typing errors evidently permits alternative explanations of the data to compete better. Even reasonably strong priors may not counteract this tendency. Fixing error rates is probably a reasonable compromise between ignoring them and estimating them from the data.

Gel 1 manifests many more obligate breaks than the combined data (see the last columns of Tables 2 and 4). Most of this excess is doubtless a consequence of typing errors. Table 5 demonstrates how these typing errors, if uncorrected, lead to inflated estimates of breakage probabilities for the best order. The total map length for the combined data is 313 cR. Under the corresponding model F for gel 1, the total map length expands to 352 cR. If typing errors are permitted, then the total map length tends to contract. This is particularly evident under model B. This model has so many error parameters with no prior con-

Table 4. Posterior Probabilities for the gel 1 data

ML rank of order	Model						Breaks
	A	B	C	D	E	F	
1	0.37336	0.16083	0.37337	0.16123	0.83293	0.94106	106
2	0.33904	0.16659	0.33904	0.16633	0.16355	0.05607	108
3	0.05563	0.08110	0.05563	0.08092	0.00235	0.00236	110
4	0.00028	0.00234	0.00028	0.00232	0.00025	0.00021	120
5	0.00028	0.00235	0.00028	0.00233	0.00012	0.00008	112
6	0.00001	0.00000	0.00001	0.00000	0.00002	0.00002	120
7	0.05485	0.08380	0.05485	0.08360	0.00046	0.00014	112
8	0.02953	0.13404	0.02953	0.13454	0.00013	0.00004	118
9	0.0000	0.00007	0.00000	0.00007	0.00001	0.00001	113
10	0.00004	0.08389	0.00004	0.08352	0.00000	0.00000	112

Models A–F are explained in the text.

Table 5. Estimated total map length, retention probability, and error rates for the gel 1 data under the best order

	Model					
Estimates	A	B	C	D	E	F
Map length (cR)	249	190	249	190	314	352
Retention	0.145	0.156	0.150	0.156	0.148	0.150
Contamination	0.015	0.017	0.008	0.016	0.005	0.000
Amplification	0.981	0.929	0.959	0.929	0.995	1.00

Averages of the contamination and amplification parameters are given for models B and D. The total map length was estimated as 313 and the retention probability as 0.150 for the combined data. Models A–F are explained in the text.

straint that we witness a drop in the estimated breakage probability θ_9 from 0.108 for the combined data to 0.019 under model B. This anomaly suggests that either fairly strong priors should be imposed on the error rates or that models with only a single amplification rate and a single contamination rate should be used.

Finally, Table 6 compares the observed and predicted proportions of typing discordancies recorded for each locus in the double typing scheme. Averaged over all loci, the observed discordancy rate is ~1%. At locus k the predicted discordancy rate is the probability

$$\sum_{g=0}^{c} \binom{c}{g} r^g (1-r)^{c-g} 2\phi_k(0 \mid g)\phi_k(1 \mid g)$$

that the two typings disagree, assuming a common retention probability r. Matching the observed average discordancy rate of 1% dictated our choice of the amplification rate of 0.995 and the contamination rate of 0.005 used in model E. It is obvious from Table 6, which is based on the model B parameter values, that estimation of amplification and contamination rates from the data yields much higher predicted discordancy rates than those observed. Even more disturbing is the almost total lack of correlation between the observed and predicted discordancy rates on a locus-by-locus basis. The first discrepancy can be explained if there is a strong correlation between the two typing results for each marker in each clone. The second discrepancy suggests that some qualitative feature of the error model is wrong. It is noteworthy that the estimation procedures had no chance to account directly for the observed discordancies between the two gels because the analyses were done one gel at a time or on the combined data ignoring typing errors.

DISCUSSION

Whole-genome, diploid radiation hybrids enjoy a decisive advantage over haploid radiation hybrids since diploid hybrids rely on a single panel to map all human chromosomes. Our theoretical development shows that diploid, and more generally polyploid hybrids, create no insurmountable computational or statistical barriers. The methods applicable to haploid hybrids carry over without major distortion to polyploid hybrids.

Table 6. Marker typing discordancy rates

Marker	Observed rate	Predicted rate	Marker	Observed rate	Predicted rate
1	0.0	0.130	8	0.0	0.095
2	0.0	0.027	9	0.012	0.071
3	0.0	0.021	10	0.0	0.0
4	0.024	0.0	11	0.024	0.061
5	0.012	0.0	12	0.012	0.023
6	0.0	0.043	13	0.048	0.104
7	0.0	0.121	14	0.012	0.073

Predicted discordancy rates are calculated under model B parameter estimates for gel 1. The average observed rate is 0.010 and the average predicted rate is 0.055.

Computing times for maximum likelihood estimation do increase but by a manageable factor of only two to five for diploid data versus haploid data. Our current optimization methods for computing posterior probabilities are faster by orders of magnitude than the recursive and Monte Carlo methods introduced previously in Lange and Boehnke (1992). Furthermore, it is not entirely clear how to generalize the earlier haploid-specific techniques to polyploid hybrids.

Our experience analyzing diploid data on 14 chromosome 4 markers prompts us to draw several tentative conclusions. First, the three optimization methods of approximating posterior probabilities for locus order appear to yield very similar answers. This suggests that the well-known device of equating posterior odds to maximum likelihood ratios is probably adequate for practical purposes (Rogatko and Zacks 1993). Second, it should come as no surprise that uncorrected typing errors cause an increase in apparent obligate breaks and a corresponding inflation of total map length. Estimating error rates has the opposite effect of deflating total map length because large error rates provide plausible alternative explanations for obligate breaks. However, there is a danger of overparameterization. Estimating too many error rates or imposing weak priors on them can degrade the posterior probability of the best order. Imposing small, but fixed error rates is probably a good compromise. Even better would be for geneticists to double type all clones. Although our analysis indicates that the two outcomes of double typing are correlated, double typing is a good strategy for ensuring high quality mapping. Lunetta et al. (this issue) compare this strategy to single typing twice as many clones. In any case, good statistical methods can only partially compensate for bad data.

Because of the limitations of space, we have not dwelt on important issues of experimental design or on strategies for locus ordering. Lange and Boehnke (1992), Chernoff (1993), and Lunetta and Boehnke (1994) consider experimental design problems relevant to haploid hybrids. Our current companion paper (Lunetta et al., this issue) extends some of this work to polyploid hybrids. Boehnke et al. (1991) outline several strategies for locus ordering. Prompted by the current plans for the construction of dense polyploid radiation hybrid maps involving hundreds of markers per chromosome, it is our intention to revisit these ordering questions. There also is the broader issue of integrating radiation hybrid maps with other types of genetic and physical maps. As the number of marker loci mapped by various methods increases exponentially, good software needs to be written to manage the enormous data processing load. Creating this software will take a careful dissection of the problems and close attention to algorithm development (Matise et al. 1995). Exploiting the full potential of polyploid radiation hybrids should remain a challenge for some time to come.

APPENDIX

Let us demonstrate strong consistency of the minimum obligate breaks criterion under a common retention probability r. Consider m loci taken in their natural order $1, \ldots, m$ along a chromosome, and imagine an infinite number of independent, fully typed radiation hybrid clones at these loci. Let $B_i(\sigma)$ be the random number of obligate breaks scored in the ith clone when the loci are ordered according to the permutation σ. In general, a permutation can be represented as an m-vector $[\sigma(1), \ldots, \sigma(m)]$. Ambiguity about the left-to-right orientation of the loci can be avoided by considering only those permutations σ with $\sigma(1) < \sigma(m)$. The correct order is given by the identity permutation $id(k) = k$.

Given h clones, the best order is identified by the permutation giving the smallest sum

$$S_h(\sigma) = \sum_{i=1}^{h} B_i(\sigma).$$

Consistency requires that $S_h(id)$ be the smallest sum for h large enough. Now the strong law of large numbers guarantees that $\lim_{h\to\infty} (1/h)S_h(\sigma) = E[B_1(\sigma)]$ with probability 1. Thus to demonstrate consistency, it suffices to show that the expected number of breaks $E[B_1(id)]$ under the identity permutation id is strictly smaller than the expected number of breaks $E[B_1(\sigma)]$ under any other permutation σ.

To compute $E[B_1(id)]$, note that the interval separating the typed loci k and $k + 1$ manifests an obligate break if and only if all of the markers at locus k are lost and at least one of the markers at locus $k + 1$ is retained, or vice versa. Given a common retention probability r, these two events are equally likely and occur with total probability $g(\theta_k) = 2(1 - r)^c[1 - (1 - \theta_k r)^c]$. Since expectations add,

$$E[B_1(id)] = \sum_{k=1}^{m-1} g(\theta_k). \qquad (24)$$

The corresponding expression for an arbitrary permutation σ is

$$E[B_1(\sigma)] = \sum_{k=1}^{m-1} g(\theta_{\sigma(k),\sigma(k+1)}), \qquad (25)$$

where $\theta_{\sigma(k),\sigma(k+1)}$ is the breakage probability for the interval $I_{\sigma(k),\sigma(k+1)}$ defined by the pair of loci $\{\sigma(k),\sigma(k+1)\}$. Obviously, $I_{\sigma(k),\sigma(k+1)}$ is a union of adjacent intervals from the correct order $1, \ldots, m$. It is plausible to conjecture that we can match in a one-to-one fashion each interval $(k,k+1)$ against a union $I_{\sigma(j),\sigma(j+1)}$ containing it. If this conjecture is true, then either $\theta_k = \theta_{\sigma(j),\sigma(j+1)}$ when the union $I_{\sigma(j),\sigma(j+1)}$ contains a single interval, or $\theta_k < \theta_{\sigma(j),\sigma(j+1)}$ when the union $I_{\sigma(j),\sigma(j+1)}$ contains several intervals. If the former case holds for all intervals $(k,k+1)$, then $\sigma = id$. If $\sigma \neq id$, the inequality $E[B_1(id)] < E[B_1(\sigma)]$ follows by taking the indicated sums (expressions 24 and 25) and noting that the function $g(\theta) = 2(1-r)^c[1-(1-\theta r)^c]$ is strictly increasing in θ for $0 < r < 1$ fixed.

Thus, the crux of the proof reduces to showing that it is possible to match one-to-one each of the intervals $(k,k+1)$ against a union set $I_{\sigma(j),\sigma(j+1)}$ that contains or covers it. This assertion is a special case of Hall's marriage theorem (Brualdi 1977). A simple direct proof avoiding appeal to Hall's theorem can be given by induction on m. The assertion is certainly true for $m = 2$. Suppose it is true for $m - 1 \geq 2$ and any permutation. There are two cases to consider.

In the first case, the last locus m is internal to the given permutation σ in the sense that σ equals $[\sigma(1), \ldots, i,m,j, \ldots, \sigma(m)]$. Omitting m from σ gives a permutation ω of $1, \ldots, m-1$ for which by induction the intervals
$(1,2), \ldots, (m-2, m-1)$
can be matched. Assuming $j < i$, the pair $\{i,j\}$ in ω covers one of the intervals $(j,j+1), \ldots, (i-1, i)$ in this matching. In the permutation σ, match the pair $\{j,m\}$ to this covered interval. This is possible because $j < i$. To the pair $\{i,m\}$ in σ match the interval $(m-1,m)$. The full matching for σ is constructed by appending these two matches to the matches for ω minus the match for the pair $\{i,j\}$. The situation with $i < j$ is handled similarly.

In the second case, m is positioned at the end of σ. By our convention this means $\sigma = [\sigma(1), \ldots, \sigma(m-1), m]$. By induction a matching can be constructed between $\omega = [\sigma(1), \ldots, \sigma(m-1)]$ and the intervals $(1,2), \ldots, (m-2, m-1)$. To this matching append the permitted match between the pair $[\sigma(m-1), m]$ and $(m-1, m)$. This completes the proof.

ACKNOWLEDGMENTS

Research was supported in part by U.S. Public Health Service grants CA16042 and HG53275 (to K.L.), HG00376 (to M.B.), and HG00206 (to D.R.C.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Barndorff-Nielsen, O.E. and D.R. Cox. 1989. *Asymptotic techniques for use in statistics*, pp. 58–68, 167–173. Chapman and Hall, London, UK.

Barrett, J.H. 1992. Genetic mapping based on radiation hybrid data. *Genomics* **13:** 95–103.

Baum, L.E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3:** 1–8.

Bishop, D.T. and G.P. Crockford. 1992. Comparisons of radiation hybrid mapping and linkage mapping. *Cytogenet. Cell Genet.* **59:** 93–95.

Boehnke, M. 1992. Radiation hybrid mapping by minimization of the number of obligate chromosome breaks. *Cytogenet. Cell Genet.* **59:** 96–98.

Boehnke, M., K. Lange, and D.R. Cox. 1991. Statistical methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.* **49:** 1174–1188.

Brualdi, R.A. 1977. *Introductory combinatorics*, pp. 153–157. North-Holland, New York.

Ceccherini, I., G. Romeo, S. Lawrence, M.H. Breuning, P.C. Harris, H. Himmelbauer, A.M. Frischauf, G.R. Sutherland, G.G. Germino, S.T. Reeders, and N.E. Morton. 1992. Construction of a map of human chromosome 16 by using radiation hybrids. *Proc. Natl. Acad. Sci.* **89:** 104–108.

Chernoff, H. 1993. Kullback-Leibler information for ordering genes using sperm typing and radiation hybrid mapping. *Statistical sciences and data analysis, proceedings of the Third Pacific Area Statistical Conference* (ed. K. Matusita, M.L. Puri, and T. Hayakawa), pp. 1–11. VSP, Utrecht, The Netherlands.

Cox, D.R., M. Burmeister, E.R. Price, S. Kim, and R.M. Myers. 1990. Radiation hybrid mapping: A somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* **250:** 245–250.

de Bruijn, N.G. 1981. *Asymptotic methods in analysis*, pp. 60–72. Dover, New York.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc. B* **39:** 1–22.

Devijver, P.A. 1985. Baum's forward-backward algorithm revisited. *Pattern Recognition Lett.* **3:** 369–373.

Feller, W. 1971. *An introduction to probability theory and its applications*, Volume 2, 2nd ed., p. 42. Wiley, New York.

Gorski, J.L., M. Boehnke, E.L. Reyner, and E.N. Burright. 1992. A radiation hybrid map of the proximal short arm of the human X chromosome spanning Incontinentia Pigmenti 1 (IP1) translocation breakpoints. *Genomics* **14:** 657–665.

Goss, S.J. and H. Harris. 1975. New method for mapping genes in human chromosomes. *Nature* **255:** 680–684.

———. 1977. Gene transfer by means of cell fusion II. the mapping of 8 loci on human chromosome 1 by statistical analysis of gene assortment in somatic cell hybrids. *J. Cell Sci.* **25:** 39–57.

Lange, K. 1995. A gradient algorithm locally equivalent to the EM algorithm. *J. R. Stat. Soc. B* **57:** 425–437.

Lange, K. and M. Boehnke. 1992. Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. *Ann. Hum. Genet.* **56:** 119–144.

Lazzeroni, L.C., N. Arnheim, K. Schmitt, and K. Lange. 1994. Multipoint mapping calculations for sperm-typing data. *Am. J. Hum. Genet.* **55:** 431–436.

Lunetta, K.L. and M. Boehnke. 1994. Multipoint radiation hybrid mapping: Comparison of methods, sample size requirements, and optimal study characteristics. *Genomics* **21:** 92–103.

Lunetta, K.L., M. Boehnke, K. Lange, and D.R. Cox. 1995. Experimental design and error detection for polyploid radiation hybrid mapping. (this issue).

Matise, T.C., C.K. Farrell, and A. Chakravarti. 1995. Automated construction of radiation hybrid maps using MultiMap. In *Genome mapping and sequencing* (ed. D. Bentley, E. Green, and R. Waterson), p. 95. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Miller, K.S. 1987. *Some eclectic matrix theory*, pp. 11–15. Robert Krieger Publishing, Malabar, FL.

Nijenhuis, A. and H.S. Wilf 1978. *Combinatorial algorithms*, pp. 240–246. Academic Press, New York.

Olson, J.M. and M. Boehnke. 1990. Monte Carlo comparison of preliminary methods for ordering multiple genetic loci. *Am. J. Hum. Genet.* **47:** 470–482.

Press, W.H., Flannery, B.P., S.A. Teukolsky, and W.T. Vetterling. 1992. *Numerical recipes. The art of scientific computing*, 2nd ed., pp. 436–443. Cambridge University Press, Cambridge, UK.

Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. Inst. Electr. Electron. Eng.* **77:** 257–285.

Rogatko, A. and S. Zacks. 1993. Ordering genes: Controlling the decision-error probabilities. *Am. J. Hum. Genet.* **52:** 947–957.

Speed, T.P., M.S. McPeek, and S.N. Evans. 1992. Robustness of the no-interference model for ordering genetic loci. *Proc. Natl. Acad. Sci.* **89:** 3103–3106.

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

Weeks, D.E. and K. Lange. 1989. Trials, tribulations, and triumphs of the EM algorithm in pedigree analysis. *IMA J. Math. Appl. Biol. Med.* **6:** 209-232.

Weeks, D.E., T. Lehner, and J. Ott. 1992. Preliminary ranking procedures for multilocus ordering based on radiation hybrid data. *Cytogenet. Cell Genet.* **59:** 125–127.

Received May 31, 1995; accepted in revised form August 16, 1995.