Experimental Design and Error Detection for Polyploid Radiation Hybrid Mapping

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

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

In this paper we consider issues of experimental design and error detection and correction for polyploid radiation hybrid mapping. Using analytic methods and computer simulation, we first consider the combinations of fragment retention rate, ploidy, and marker spacing that provide the best chance to order markers. We find that in general, combinations of ploidy and chromosome-specific retention rates that lead to a per-hybrid retention rate of ~50% result in the greatest power to order markers. We also find that analyzing polyploid radiation hybrids as if they were haploid does not compromise the ability to order markers but does result in less accurate intermarker distance estimates. Second, we examine the effect of typing errors on two-locus information, ability to order multiple loci, and estimation of intermarker distances and total map length. Even low levels of error result in large losses of information about breakage probabilities, markedly reduce ability to order loci, and inflate estimates of intermarker distances and total map length. We compare the ordering accuracy that results from duplicate typing of hybrids to that of single typing twice as many hybrids and find that duplicate typing results in a higher probability of identifying the true order as one of the best orders, but that single typing of twice as many hybrids results in stronger support for the true order. For low error rates, framework maps constructed from the larger single-typed panels are only slightly less likely to be correct and include substantially more markers than the smaller double-typed panels. Third, we develop a method to calculate the distribution of the number of obligate chromosome breaks for a polyploid radiation hybrid under a given locus order and discuss how this method may be used to identify hybrids with suspiciously large numbers of chromosome breaks.

Recently, investigators have turned to whole-genome diploid radiation hybrids (RHs) as a tool for ordering and estimating distances between loci on any human chromosome (Walter et al. 1994; D.R. Cox, R.M. Myers, D. Vollrath, M. Boehnke, and K. Lange, in prep.). This shift away from the methods developed by Cox et al. (1990), which use human–rodent hybrids containing a single human chromosome, represents a return to the ideas originally developed by Goss and Harris (1975, 1977a,b).

In the Goss and Harris formulation, normal diploid human cells were lethally irradiated, breaking each human chromosome into several fragments. The irradiated cells were then fused with a hamster cell line deficient in the enzyme hypoxanthine phosphoribosyl transferase (HPRT). Growth in HAT medium selected those fused cells containing the human HPRT gene, lo-cated on the human X chromosome. Loci on the X chromosome were then mapped. The human–hamster hybrids tended to retain only human chromosome fragments that contained the selected locus (Goss and Harris 1975, 1977a), limiting the usefulness of the procedure to mapping genes located near the human HPRT gene or other selectable loci. Later, Goss and Harris (1977b) switched to human–mouse hybrids to order loci on human chromosome 1. These hybrids appeared to retain at random chromosome fragments that do not contain the selected locus.

Cox and colleagues (Cox et al. 1990; Burmeister et al. 1991) simplified the methods of Goss and Harris (1977b) by using a Chinese hamster–human somatic cell hybrid containing only a single copy of a human chromosome as the irradiated donor cell line. The irradiated cells containing the fragmented human chromosomes were rescued by fusion with a normal HPRT-deficient rodent cell line and grown in HAT me-

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

dium, so that only those donor–recipient hybrids that contained the hamster HPRT gene from the irradiated donor cell survived. This haploid version of RH mapping has been used successfully to map markers on many human chromosomes (e.g., 8, Oshima et al. 1994; 11, Richard et al. 1991, 1993; 17, Abel et al. 1993; 21, Cox et al. 1990; Burmeister et al. 1991; 22, Frazer et al. 1992; X, Gorski et al. 1992). As in the original Goss and Harris (1977b) formulation, the patterns of presence and absence of various markers in the hybrid clones are used to infer marker order by exploiting the principle that the closer together two markers are on a chromosome, the fewer radiation-induced breaks occur between them. A variety of strategies for ordering markers and estimating intermarker distances for haploid RH data have been developed (Cox et al. 1990; Boehnke et al. 1991; Falk 1991; Chakravarti and Reefer 1992; Lange and Boehnke 1992; Weeks et al. 1992; Wilson 1992).

Haploid RH mapping has the advantage that in the absence of typing error, the number of copies of each human marker in each hybrid clone is observable. If a clone tests positive, one copy of the marker is present; if it tests negative, zero copies are present. The primary disadvantage of the haploid approach is that it is labor intensive: A separate panel of hybrids must be constructed to map each human chromosome. To address this problem, Walter et al. (1994) and D.R. Cox, R.M. Meyers, D. Vollrath, M. Boehnke, and K. Lange (in prep.) have developed a whole-genome RH mapping approach, similar to the original Goss and Harris (1977b) procedure. This revised procedure involves irradiating a diploid human cell line rather than a rodent–human hybrid cell line. The advantage of the whole-genome approach is that a single set of hybrids may be used to map all human chromosomes. A disadvantage of diploid hybrids is that we can tell only whether a marker is present or absent in a hybrid. If a marker is present, we do not know whether a single copy or two copies are present. Further, preliminary results for some diploid hybrids suggest that the human chromosomal fragments may be retained at a lower rate in diploid hybrids than in haploid hybrids (D.R. Cox, R.M. Meyers, D. Vollrath, M. Boehnke, and K. Lange, in prep.). For example, the chromosome-specific retention rate for the diploid hybrids distributed by Research Genetics is ~8% per chromosome, or 16% per diploid hybrid (D.R. Cox, R.M. Meyers, D. Vollrath, M. Boehnke, and K. Lange, in prep.).

Lower hybrid-specific retention rates result in less information for ordering markers. This problem can be overcome by pooling n diploid hybrids to achieve $2n$-ploid hybrids with per-hybrid retention close to some optimal value that provides the maximum probability of correct marker ordering.

Lange et al. (1995) describe computational methods for the statistical analysis of diploid and polyploid radiation hybrids. In this paper we consider issues of optimal design involving ploidy, fragment retention rates, and typing errors. We also develop a method to calculate the distribution of the number of obligate chromosome breaks for diploid and more generally polyploid radiation hybrids, and discuss the use of the distribution as a tool for detecting hybrids exhibiting an excessive number of obligate breaks.

Notation and Assumptions

Suppose markers $A_1, \ldots, A_M$ are typed on H radiation hybrids. The observations for a hybrid given the marker order $(A_1, \ldots, A_M)$ can be written as $x = (x_1, \ldots, x_M)$, where $x_i = 1$ if the marker is typed and present, 0 if the marker is typed and absent, and ? if the marker was not typed or if two or more independent typings result in conflicting scores. We define $c \geq 1$ to be the ploidy of the panel. For haploid and diploid panels, $c = 1$ and $c = 2$ respectively. For polyploid panels, $c \geq 2$. The vector $g = (g_1, \ldots, g_M)$ defines the actual number of copies of each marker retained in a single hybrid, where $0 \leq g_i \leq c$.

We define the breakage probability θ_i as the probability of at least one chromosome break between markers A_i and A_{i+1}, and the distance d_i as the expected number of breaks per hybrid between these markers. We assume that breakage occurs at random along the chromosome, so that breakage for a single chromosome can be modeled as a Poisson process. Under this assumption the breakage probability θ and the distance d are related as $d = \log_e(1 - \theta)$, in analogy to Haldane's (1919) no interference mapping function for linkage analysis. The distance d is measured in Rays, where one Ray is equivalent to one expected break per hybrid.

We define the per-chromosome retention probability r_{ij} $(i \leq j)$ to be the probability that a fragment spanning markers $A_i, \ldots, A_j$ is retained in a hybrid. Breakage and retention are considered independent processes. For ease of presentation throughout this paper, we assume $r_{ij} = r_i$ for $1 \leq i,j \leq M$ (Boehnke et al. 1991). That

is, r_i is the probability that a fragment with left-most marker A_i will be retained in a hybrid. As special cases of this left-endpoint model, we have the equal retention model in which all fragments have the same retention probability r, and the centromeric retention model in which fragments containing the marker nearest the centromere have a different retention probability than all other fragments. We assume an equal retention model for all of our simulations. For that model, the per-hybrid rate of marker retention for a c-ploid hybrid is $r_o = 1 - (1 - r)^c$. We use 500 replicate data sets for all simulations.

In our analyses of typing error, we define the false-positive rate π_i to be the probability that marker A_i is scored as present in the hybrid when it is absent from the hybrid. The false-negative rate $v_i(g_i)$ is the probability that marker A_i is scored as absent given that there are $g_i \geq 1$ copies present in the hybrid. In our simulations and analyses we assume that the false-negative rate does not depend on the number of copies of the marker present in the hybrid [$v_i(g_i) = v_i \geq 1$], and that the error rates for all markers are the same. That is, $\pi_i = \pi$ and $v_i = v$ for $i = 1, \ldots, M$. We assume further in some of our simulations and analyses that the false-positive and false-negative rates are equal ($\pi = v$) and denote the common error rate by ϵ.

Optimal Design

We used two approaches for exploring optimal design. First, we calculated the Fisher information about the breakage probability θ for the two-locus case and determined the ploidy c and retention probability r that result in maximum information for a given value of θ. Second, we evaluated by computer simulation the probability of correctly ordering multiple markers and estimated values of c and r which allowed us to order the loci most accurately. Finally, we examined the strategy of analyzing polyploid ($c > 1$) data assuming a haploid model ($c = 1$).

Information About the Breakage Probability

The Fisher information for the breakage probability θ between two markers can be expressed in terms of the coretention probabilities

$$q_{00} = s^c(1 - \theta r)^c$$

$$q_{01} = s^c[1 - (1 - \theta r)^c] \qquad = q_{10}$$

$$q_{11} = 1 - s^c[2 - (1 - \theta r)^c]$$

(1)

derived in Lange et al. (1995), where $s = 1 - r$. In this multinomial setting, the Fisher information $I_{\theta\theta}$ for a single observation is given by

$$I_{\theta\theta} = \sum_{i=0}^{1} \sum_{j=0}^{1} \left(\frac{\partial q_{ij}}{\partial \theta}\right)^2 \frac{1}{q_{ij}}.$$

(2)

Slightly tedious algebra that we omit shows that expressions 1 and 2 imply

$$I_{\theta\theta} = \frac{c^2 r^2 s^c (1 - \theta r)^{c-2}[1 - 2s^c + (1 - \theta r)^c]}{[1 - (1 - \theta r)^c][1 - 2s^c + s^c(1 - \theta r)^c]}.$$

Table 1 displays the ploidy c resulting in maximal information about a breakage probability θ and the range of ploidies (c_l, c_u) resulting in at least 90% of the maximum information as a function of r and θ. For example, given retention $r = 0.08$, optimal ploidy by this criterion ranges from $c = 11$ for $\theta = 0.10$ to $c = 8$ for $\theta = 0.40$. The column (r_{ol}, r_{ou}), where $r_{ol} = 1 - (1 - r)^{c_l}$ and $r_{ou} = 1 - (1 - r)^{c_u}$ shows the range of per-hybrid retention rates that correspond to r and (c_l, c_u). In all cases, the range for r_0 includes 0.50, which is the retention probability that results in maximum information in the haploid case (Lange and Boehnke 1992).

Table 1. Ploidy (c) yielding the maximum information about a breakage probability θ given per-chromsome retention probability r; range of ploidies (c_l, c_u) and per-hybrid retention probability range (r_{ol}, r_{ou}) giving $\geq$ 90% of the maximum information

r	θ	c	(c_l, c_u)	(r_{ol}, r_{ou})
0.08	0.10	11	(7,16)	(0.44,0.74)
	0.20	10	(6,14)	(0.39,0.69)
	0.30	9	(6,13)	(0.39,0.66)
	0.40	8	(5,12)	(0.34,0.63)
0.10	0.10	9	(6,13)	(0.47,0.75)
	0.20	8	(5,11)	(0.41,0.69)
	0.30	7	(5,10)	(0.41,0.65)
	0.40	6	(4,9)	(0.34,0.61)
0.12	0.10	7	(5,10)	(0.47,0.72)
	0.20	6	(4,9)	(0.40,0.68)
	0.30	6	(4,8)	(0.40,0.64)
	0.40	5	(4,8)	(0.40,0.64)
0.25	0.10	3	(2,4)	(0.44,0.68)
	0.20	3	(2,4)	(0.44,0.68)
	0.30	3	(2,4)	(0.44,0.68)
	0.40	2	(2,3)	(0.44,0.58)

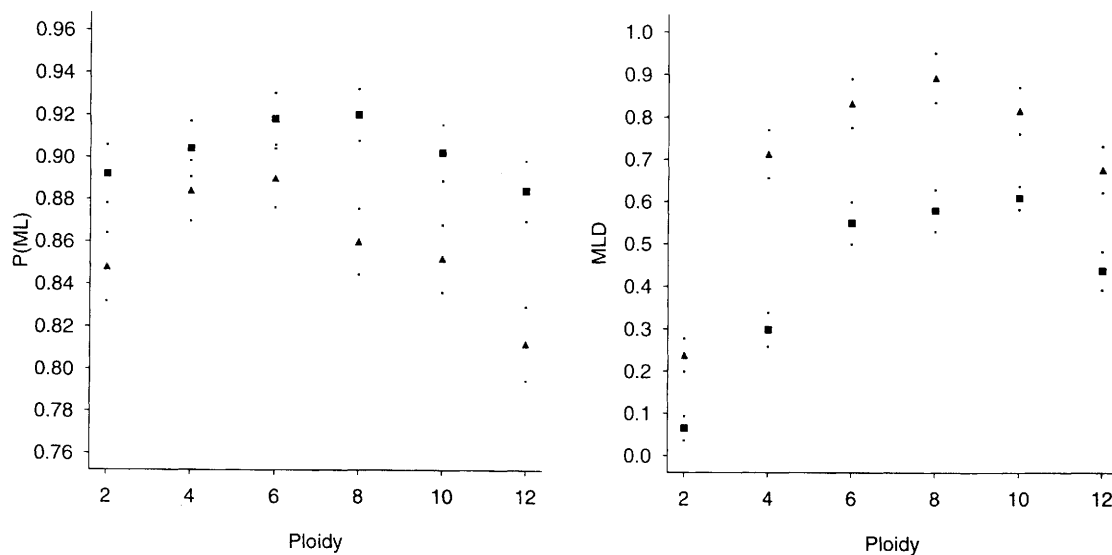*Ordering Accuracy for Multiple Markers*

We simulated data with $H = 100$ hybrids and $M = 8$ markers randomly distributed along an interval of length $d \times (M + 1)$ Rays, where $d = 0.1$ and 0.2. This results in average intermarker distances of 0.1 and 0.2 Rays and an expected map length of $E(L) = (M - 1) \times d$ Rays. We assumed fragment retention probability $r = 0.08$, as observed in the diploid high-dose radiation whole-genome hybrid panel of D.R. Cox, R.M. Meyers, D. Vollrath, M. Boehnke, and K. Lange (in prep.) and used maximum likelihood to order the loci (Boehnke et al. 1991; Lange et al. 1995). To compare the ordering accuracy for sets of values of H, M, c, d, and r, we used two optimality criteria. P(ML) is the probability that the true order is a maximum likelihood order. One or more other orders may have a maximum likelihood equal to the true order maximum likelihood. MLD is the log-likelihood difference between the true and the (next) most likely order. If the true order is the maximum likelihood order, MLD is the log-likelihood difference between true order and the order with next largest maximum likelihood, and MLD ≥ 0. If the true order is not the maximum likelihood order, then MLD is the log-likelihood difference between the true order and the maximum likelihood order, and MLD < 0.

The mean MLD and estimates of P(ML) are plotted in Figure 1 for $H = 100$ hybrids, retention rate $r = 0.08$, and $M = 8$ markers for average marker spacings 0.10 and 0.20 rays. By both criteria, the range of ploidies $c = 6 - 10$ generally results in the best performances under the conditions simulated. The increase in support for the true order (MLD) is particularly strong between diploid and tetraploid ($c = 4$) hybrids. Overall, these results suggest that for $r = 0.08$, 8-ploid data results in the best ability to order the markers. The per-hybrid retention rate in this case is $r_0 = 0.49 \approx 0.50$.

Analysis Strategy

Analyses of 8-ploid data assuming $c = 8$ and assuming a haploid ($c = 1$) model are summarized in Table 2. For example, for average spacing 0.1 Ray and chromosome-specific retention $r = 0.08$, the mean maximum log-likelihood difference (MLD) for the haploid analysis was 0.598, versus 0.559 for the 8-ploid analysis. The estimated probability that the true order is a maximum likelihood order [P(ML)] is also always slightly greater for the $c = 1$ analyses than the $c = 8$ analysis. For the cases considered, haploid analysis appears to result in slightly better ability to order by the MLD and P(ML) criteria. However, the haploid analysis results in strong overestimation of the intermarker distances d_i and total map length L. We define the estimated map length as

Figure 1 Plots of two criteria (described in text) used to compare ordering accuracy of multiple loci. Data are from simulations using $H = 100$ hybrids and $M = 8$ markers randomly spaced with average intermarker distances of 0.1 and 0.2 Rays, and fragment retention probability $r = 0.08$. (■) 0.1 Ray average spacing; (▲) 0.2 Ray average spacing; (dots) ± 1 s.e.

Table 2. Estimated P (ML) and mean estimates of MLD and $B(L)$ for the true order: 8-ploid data analyzed assuming the correct ploidy ($c = 8$) and assuming hybrids are haploid ($c = 1$). Simulated data: $H = 100$, $M = 8$, $c = 8$, average marker spacing d

d	r	P(ML)		MLD		$B(\hat{L})$	
		$c = 8$	$c = 1$	$c = 8$	$c = 1$	$c = 8$	$c = 1$
0.1	0.08	0.892	0.896	0.559	0.598	0.062	0.364
	0.12	0.888	0.898	0.530	0.588	0.048	0.527
	0.20	0.816	0.820	0.158	0.191	0.073	0.979
0.2	0.08	0.878	0.884	0.917	0.986	0.063	0.410
	0.12	0.838	0.854	0.614	0.705	0.065	0.505
	0.20	0.712	0.696	0.171	0.196	0.129	0.982

S.E.: ≤ 0.01[P(ML)]; 0.036–0.062 (MLD); and 0.014–0.032 [$B(\hat{L})$].

Fisher information about θ as a function of the error rates and compared this information to the information when the error rates were zero. Second, we examined the asymptotic proportional bias in estimating θ to determine the effect of typing error on the estimated map length. Third, we simulated data with error and analyzed it assuming no error or allowing for error to determine the effects on ordering accuracy and estimated map length. Fourth, we examined the effect of duplicate typing of RH data on ability to order markers, estimate model parameters, and place markers in framework maps.

$$\hat{L} = \sum_{i=1}^{M-1} \hat{d}_i = \sum_{i=1}^{M-1} \log_e(1 - \hat{\theta}_i)$$

and use the proportional bias $B(\hat{L}) = (\hat{L} - L)/L$ as a measure of map expansion. Under the conditions that we simulated, $B(\hat{L})$ generally is small for the $c = 8$ analysis with upward bias in map length of 5%–13%. In contrast, for the $c = 1$ analysis, upward bias in map length reached nearly 100% (Table 2). The mean differences in maximum loglikelihood for the true order between the $c = 8$ and $c = 1$ analyses are small, ranging from 0.046 to 0.223 (data not shown), suggesting that the incorrect haploid model fit the data nearly as well as the correct 8-ploid model.

Analysis using $c > 1$ is more computationally demanding than haploid ($c = 1$) analysis. Analyses assuming $c = 2$ required between 2 and 5 times longer than analyses assuming $c = 1$; analyses assuming $c = 8$ required 14 to 20 times longer. Given the equivalent, or perhaps even superior ordering ability for haploid analysis and the differences in analysis times, a plausible strategy for analysis of polyploid data is to first analyze the data assuming $c = 1$ to obtain a set of best orders and then to reanalyze the set of best orders using the correct ploidy to confirm the best order and to obtain more accurate intermarker distance estimates.

Effect of Typing Errors

To examine the effect of typing errors we again used a combination of analytic calculation and computer simulation. First, we calculated the

Relative Information with Typing Error

To compute the Fisher information about the breakage probability θ as a function of the false-positive rate π and false-negative rate v, we first compute the coretention probabilities $q_{ij}^* = P(x = (i,j)|\pi,v)$. Recall that $q_{ij} = P(x = (i,j)|\pi = v = 0)$ are the coretention probabilities assuming no error defined in expression 1. Then assuming independent errors,

$$q_{00}^* = (1 - \pi)^2 q_{00} + 2(1 - \pi)v q_{01} + v^2 q_{11}$$

$$q_{01}^* = \pi(1 - \pi)q_{00} + [(1 - \pi)(1 - v) + \pi v]q_{01} \qquad (3)$$
$$+ v(1 - v)q_{11} = q_{10}^*$$

$$q_{11}^* = \pi^2 q_{00} + 2\pi(1 - v)q_{01} + (1 - v)^2 q_{11}.$$

Replacing q_{ij} in expression 2 by q_{ij}^*, we compute the information $I_{\theta\theta}(\pi,v)$ about θ as a function of π and v. Figure 2 displays the relative information about θ, $R_\theta(\pi,v) = I_{\theta\theta}(\pi,v)/I_{\theta\theta}(0,0)$ for ploidy $c = 2$ and 8, as a function of the equal false-positive and false-negative rates ($\pi = v = \epsilon$) for retention probability $r = 0.08$. Overall, the information loss due to typing error is greatest for closely spaced markers and low ploidy. With error rate $\epsilon = 0.005$

and $\theta = 0.1$, typing error results in an ~30% loss of information about θ when $c = 2$ and ~15% when $c = 8$. With error rate $\epsilon = 0.02$ and $\theta = 0.1$, typing error results in an ~60% loss of information about θ when $c = 2$ and ~40% when $c = 8$.

Asymptotic Proportional Bias

For two markers, we are able to examine analytically the asymptotic proportional bias in estimating θ when we have typing error and can use this as a measure of the effect of the errors on the estimation of map length. If $\epsilon = 0$, $P(x = (0,1)) = q_{01} = s^c[1 - (1 - \theta r)^c]$, where $s = 1 - r$. Hence,

$$\theta = \frac{s - (s^c - q_{01})^{1/c}}{rs} . \tag{4}$$

Because maximum likelihood estimators are asymptotically unbiased, the maximum likelihood estimator $\hat{\theta}$ has limiting expectation given in expression 4. If the false-positive and false-negative error rates are not both zero, then $q_{01} \neq s^c[1 - (1 - \theta r)^c]$, and θ^*, the limiting expectation of $[s - (s^c - q_{01})^{1/c}]/rs$ will not be equal to θ. We define the asymptotic proportional bias as $(\theta^* - \theta)/\theta$. We display this bias for $c = 2$ and 8 in Figure 3. The asymptotic proportional bias decreases with increasing θ, and approximately doubles as the error rates double. The bias is more severe in the diploid ($c = 2$) case where error rate $\epsilon = 0.02$ and breakage probability $\theta = 0.1$ result in a proportional bias of 1.4, or an estimate of θ that is ~2.4 times the actual value of $\theta = 0.1$.

Ordering Accuracy and Bias for Multiple Markers

To investigate ordering accuracy and bias in map-length estimates for data with typing errors, we simulated data with $H = 100$ hybrids, $M = 8$ markers, equal retention $r = 0.08$, and average inter-marker spacing of 0.1 and 0.2 Rays. To these data we introduced random errors at rate $\epsilon = 0.005$ or $\epsilon = 0.01$. We then analyzed the data sets (1) under
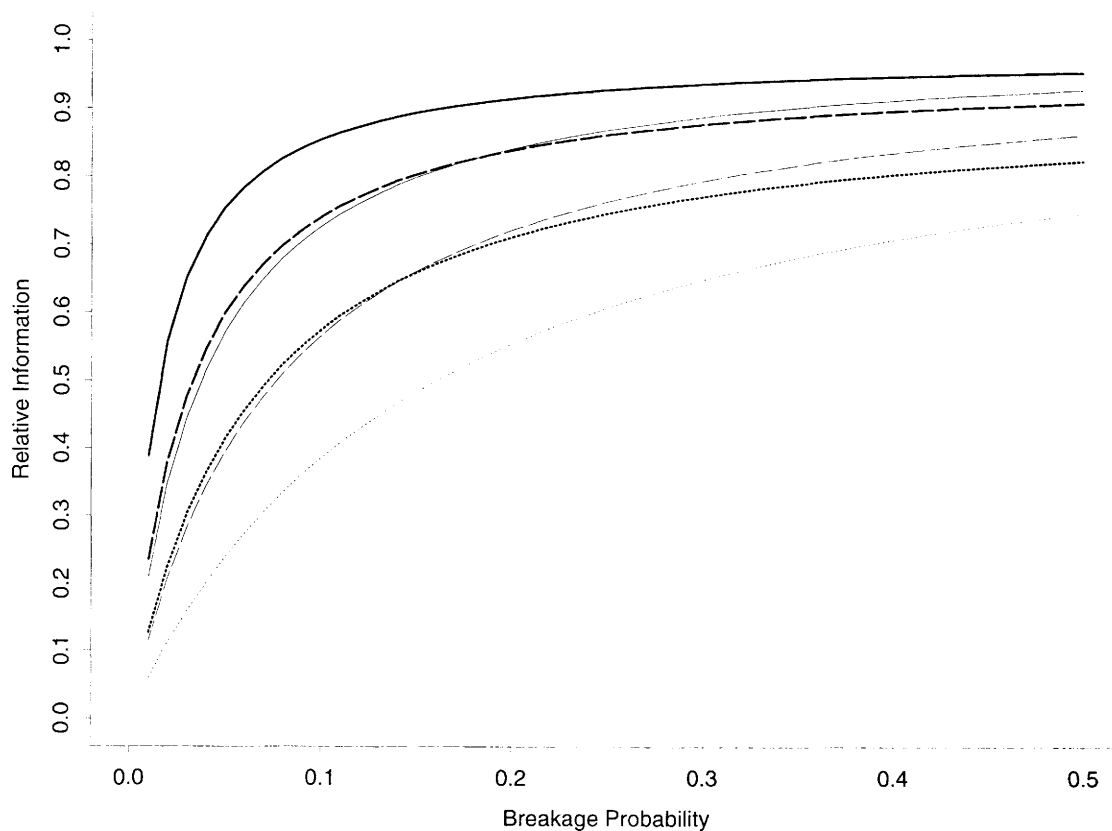
Figure 2 Relative information about the breakage probability θ assuming equal fragment retention probability $r = 0.08$ and equal error probabilities (ϵ) when data are analyzed assuming no typing error. (Light lines) $c = 2$; (dark lines) $c = 8$; (smooth lines) $\epsilon = 0.005$; (broken lines) $\epsilon = 0.01$; (dotted lines) $\epsilon = 0.02$.
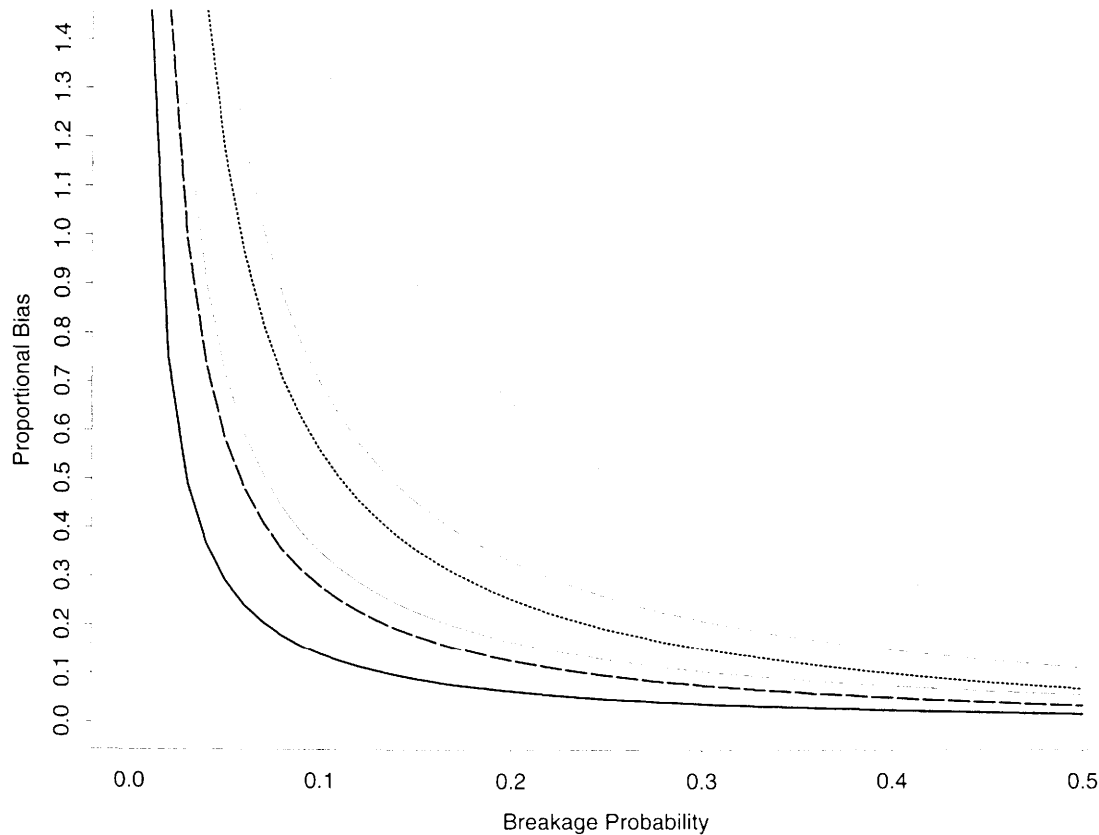
Figure 3 Asymptotic proportional bias in the estimate of the breakage probability θ assuming equal fragment retention probability $r = 0.08$ and equal error probabilities (ϵ) when data area analyzed assuming no typing error. (Light lines) $c = 2$; (dark lines) $c = 8$; (smooth lines) $\epsilon = 0.005$; (broken lines) $\epsilon = 0.01$; (dotted lines) $\epsilon = 0.02$.

the no-error assumptions $\epsilon = 0$, and (2) correctly taking into account the true error rates. For comparison, we also analyzed the data sets before error was introduced. Besides the effect on ordering accuracy, we are concerned with the degree of inflation in map-length estimates resulting from typing error; we examined the proportional bias $[B(\hat{L} = (\hat{L} - L)/L]$ as a measure of this increase.

The probability that the true order is a maximum likelihood order [P(ML)] is considerably higher for data with no typing error than for data with typing error, even when typing error is taken into account in the analysis (Table 3). The differences are more severe for the diploid data than for the 8-ploid data. For simulations with no error ($\epsilon = 0$), the mean $B(\hat{L})$ is positive but small and is probably attributable to the small numbers of hybrids (Lunetta and Boehnke 1994). Given typing error in the data, the estimates of map length for the correct order show considerable positive bias. For analyses that take into account the typing error rate, the mean proportional bias in map length is reduced greatly relative to the

analyses that ignore the typing error and is similar to that for data with no typing error. For given error rates, the proportional bias is greater for $c = 2$ than $c = 8$ and for 0.10 Ray average marker spacing than for 0.20 Ray average spacing. The proportional bias for $\epsilon = 0.01$ is somewhat smaller than twice that for $\epsilon = 0.005$. The average proportional bias in the estimates of the individual breakage probabilities, and hence in the total map length, are similar to those predicted by our two-locus analytic calculations. For example, for $c = 2$, $\epsilon = 0.01$, and average $d = 0.2$ Rays (or breakage probability $\theta \approx 0.18$), the average of the estimates for the seven breakage probabilities is 0.2314 (data not shown), so that the proportional bias in the estimates of the breakage probabilites is ~0.28. Our analytic calculation yields 0.37 (Fig. 3). Other values of c, ϵ, and d all result in smaller differences between the results from simulation and analytic calculations. The differences in our analytic and simulation results are likely to be attributable in part to the fact that the simulated markers are not equally spaced.

LUNETTA ET AL.

Table 3. Estimated P(ML) and mean estimates of $\hat{r}$ and $B(\hat{L})$ for the true order: data with typing error rates ϵ analyzed assuming no typing error and assuming the true typing error rate. Simulated data: $H = 100$, $M = 8$, retention $r = 0.08$. [P(ML), $\hat{r}$, and $B(\hat{L})$ described in text]

			Error rate assumed in analysis					
			0			ϵ		
Data				mean			mean	
c	d	ϵ	P(ML)	$\hat{r}$	$B(\hat{L})$	P(ML)	$\hat{r}$	$B(\hat{L})$
2	0.1	0.00	0.880	0.080	0.078	0.880	0.080	0.078
	0.1	0.005	0.458	0.082	0.487	0.492	0.080	0.090
	0.1	0.01	0.306	0.084	0.895	0.390	0.080	0.098
	0.2	0.00	0.842	0.080	0.059	0.842	0.080	0.059
	0.2	0.005	0.568	0.082	0.251[a]	0.568	0.080	0.064[a]
	0.2	0.01	0.458	0.084	0.447[a]	0.472	0.080	0.074[a]
8	0.1	0.00	0.898	0.081	0.025	0.898	0.081	0.025
	0.1	0.005	0.612	0.081	0.195	0.608	0.081	0.032
	0.1	0.01	0.488	0.081	0.363	0.468	0.081	0.033
	0.2	0.00	0.876	0.080	0.023	0.876	0.080	0.023
	0.2	0.005	0.704	0.080	0.102	0.702	0.080	0.022
	0.2	0.01	0.592	0.080	0.187	0.604	0.080	0.022

s.e.: ≤0.001 [P(ML)]; <0.001 ($\hat{r}$); 0.008–0.022 [$B(\hat{L})$].
[a]Excludes one replicate in which the estimate of one of the breakage parameters was 1.0.

Considering the maximum likelihood orders rather than the true orders gives similar results, although the map length tends to be slightly less biased (data not shown). This smaller bias can be explained by the fact that when it is not a maximum likelihood order, the true order tends to have more obligate chromosome breaks than the maximum likelihood orders and, therefore, larger estimates for intermarker distances d_i and total map length L.

Effect of Duplicate Typing

One method to minimize typing errors in RH data is to type the hybrid clones twice and to treat as missing ($x_i = ?$) any markers that show discordant results in the two typings. If we assume that the two typings are independent and the false-positive and false-negative error rates are ϵ, then the rate of discordance in the two typings is $2\epsilon(1 - \epsilon) \approx 2\epsilon$. That is, we expect the discordance rate to be approximately twice the typing error rate. Double typing the data greatly reduces the number of errors but also doubles the amount of work required to type each hybrid. For the same genotyping effort, a panel of twice as many hybrids could be typed once for each marker. To determine whether the single or double typing strategy is more powerful, we simulated diploid ($c = 2$) data sets of $H = 200$ hybrids with false-positive and false-negative error rates of $\epsilon = 0.0025$, 0.005, and 0.010 to correspond to single-typed data. We simulated data sets of $H = 100$ hybrids without error but with missing data at rates $2\epsilon = 0.005$, 0.01, and 0.02 to correspond to double-typed data. We analyzed the double-typed hybrids assuming $\epsilon = 0$. We analyzed the single-typed data with error assuming (1) no error, (2) one-half the true error rate, (3) the true error rate, and (4) twice the true error rate. Table 4 provides a summary of our results. For $\epsilon = 0.005$, the probability that the true order is a maximum likelihood order [P(ML)] is greater for the 100 double-typed hybrids than for the 200 single-typed hybrids under any assumed error rate. The difference is smaller for $\epsilon = 0.0025$ and more substantial for $\epsilon = 0.01$. In contrast, the MLD for the true order is greater for the 200 single-typed hybrids than for the 100 double-typed hybrids for all but the analyses assuming the error rates are twice the true values. In general, we see that the 200 single-typed hybrids allow us to identify the true order as the best order with greater support (MLD) than the 100 double-typed hybrids but that the 100 double-typed hybrids are more likely to find that the true order as one of the maximum likelihood orders. Furthermore, we note that the error rate at which we analyze the single-typed data does not strongly effect

Table 4. Estimated P(ML) and mean estimates of MLD and $B(\hat{L})$ for the true order: $H = 200$ single-typed hybrids vs. $H = 100$ double-typed hybrids. Simulated diploid data: $M = 8$, retention $r = 0.08$, average marker spacing $d = 0.20$ rays, error rates ϵ

			Error rate assumed in analysis			
				single typed		
True error rate (ϵ)		double typed[a] [0.0]	0.0	$\frac{1}{2} \times \epsilon$	ϵ	$2 \times \epsilon$
0.0025	P(ML)	0.842	0.760	0.774	0.774	0.778
	MLD	0.223	0.932	0.754	0.600	0.418
	$B(\hat{L})$	0.059	0.130	0.062	0.029	−0.015
0.005	P(ML)	0.812	0.694	0.726	0.734	0.740
	MLD	0.267	0.812	0.604	0.428	0.245
	$B(\hat{L})$	0.049	0.228	0.088	0.029	−0.038
0.010	P(ML)	0.774	0.602	0.624	0.650	0.624
	MLD	0.176	0.494	0.369	0.229	0.086
	$B(\hat{L})$	0.054	0.422	0.133	0.032	−0.071

S.E.: 0.006–0.011 [P(ML)]; 0.017–0.092 (MLD); 0.008–0.012 [$B(\hat{L})$].
[a]Double-typed data have missing observations at rate $2 \times \epsilon$.

For error rates $\epsilon \leqslant 0.01$, the larger panel of single-typed hybrids consistently allows us to place >70% more markers into a 1000:1 framework map than the smaller double-typed panel, whereas the estimated probability that a framework map is incorrect is only slightly greater. For an error rate as large as $\epsilon = 0.02$ the larger panel does not show as clear an advantage. In that case, MIM is still >70% greater for the larger single-typed panel; however, P(correct) = 0.994 for the double-typed hybrids, but only 0.924 for the single-typed hybrids.

P(ML) but influences MLD more strongly. Finally, allowing for typing error in the analysis greatly reduces the proportional bias in the estimate of the total map length, even if we assume an error rate as small as one-half the true error rate.

Because of the favorable comparison of single typing 200 hybrids to double typing 100 hybrids, we further examined the two strategies by comparing their success at building framework maps. As before, we simulated data sets of $H = 200$ hybrids with random errors and $H = 100$ hybrids with random missing data to correspond to single- and double-typed data. For this analysis, we simulated $M = 16$ markers with average spacing 0.2 ray and retention rate $r = 0.08$. Then we attempted to build maps including as many markers as possible in which all markers are placed with 1000:1 relative odds. We compared the mean number of markers that could be placed in the map (MIM) and the estimated probability that the markers in the map were placed in the correct order [P(correct)] for the two typing strategies.

For framework map building, allowing for error in the analysis have virtually no effect on either the mean number of markers placed at 1000:1 relative odds or the percentage of the maps that were correct. Therefore, in Table 5 we report the results when no error is assumed in the analysis.

Distribution of the Number of Obligate Chromosome Breaks

It is clear that typing error can have a substantial effect on accuracy of locus ordering and estimation of map length. Because typing errors tend to increase the number of obligate chromosome breaks for a given order, one way to identify individual hybrids with possible typing errors is to look for hybrids with surprisingly large numbers of obligate breaks under the best order. Such hybrids may be identified by calculating the probability of obtaining the observed number of obligate breaks or more assuming no typing error. Ideally, the distribution would be calculated under the true order using the true breakage and retention probabilities. In practice, we must use the best order identified and the parameter estimates for that order. For the entire hybrid panel, we can determine how well the observations fit our model assumptions by comparing the observed and expected numbers of hybrids with $b = 0, 1, \ldots, M - 1$ obligate breaks.

LUNETTA ET AL.

Table 5. Mean number of markers placed in a 1000:1 framework map (MIM) and estimated probability that the map is correct [P(correct)]: $H = 200$ single-typed hybrids vs. $H = 100$ double-typed hybrids. Simulated diploid data: $M = 16$ markers, retention $r = 0.08$, average marker spacing $d = 0.20$ Rays, error rates ϵ. Error rate $= 0.0$ assumed in all analyses

True error rate ϵ		Double typed[a] ($H = 100$)	Single typed ($H = 200$)
0.0025	MIM	6.742	11.936
	P(correct)	0.998	0.992
0.005	MIM	6.808	11.790
	P(correct)	1.000	0.986
0.010	MIM	6.472	11.394
	P(correct)	0.998	0.974
0.020	MIM	6.120	10.654
	P(correct)	0.994	0.924

s.e.: ≤0.001 [P(correct)]; 0.111–0.129 (MIM).
[a]Double-typed data have missing observations at rate $2 \times \epsilon$.

A straightforward recurrence relation allows us to calculate the distribution of the number of obligate breaks for a single hybrid. For simplicity, we assume initially that all markers are typed ($x_i = 0$ or 1). Let $P_n(b,k,w)$ be the joint probability of b obligate breaks among the first n loci of an order in a hybrid with k copies of locus n retained and observation $x_n = w$. Let $\phi_i(w|k) = P(x_i = w|g_i = k)$, the probability that we observe marker A_i to be present ($w = 1$) or absent ($w = 0$) given that k copies of the marker are retained in the clone. For example, $\phi_i(1|0) = \pi$ and $\phi_i(0|k) = v_i(k)$ for $k > 0$. If we think of x_i as the phenotype and g_i as the genotype of marker i for the hybrid, the probability ϕ_i is analogous to a penetrance function in linkage analysis. Consider the marker order $(A_1, A_2, \ldots, A_M)$. Given a single marker A_1, there can be no obligate breaks, and the initial conditions are

$$P_1(0,k,w) = \binom{c}{k} r_1^k (1 - r_1)^{c-k} \phi_1(w|k). \quad (5)$$

That is, the probability that k copies of the leftmost marker are retained in the hybrid and that we observe retention status w is the binomial probability that k copies of the marker are re-

tained multiplied by the probability that we observe retention status w given k copies of the marker are retained. Starting with n markers, when we type the next marker on the hybrid, we add at most one obligate break. Hence, for $n \geq 1$,

$$P_{n+1}(b,k,w) = \sum_{j=0}^{c} [P_n(b,j,w)\phi_{n+1}(w|k) + P_n(b-1,j,1-w)\phi_{n+1}(1-w|k)] \times t_{c,n}(j,k) \quad (6)$$

where we define $P_n(-1,g,w) = 0$ for $0 \leq g \leq c$, and $w = 0$ or 1, and $t_{c,n}(j,k)$ is the probability for c-ploid data that k copies of marker A_{n+1} are retained given that j copies of marker A_n are retained (see below). Because $1 - w$ is 0 when w is 1 and 1 when w is 0, the recurrence (expression 6) increments the obligate break count whenever the observations for two adjacent typed markers differ.

To obtain the transition probabilities $t_{c,n}(j,k)$, we first note that the transition probabilities for the haploid ($c = 1$) case satisfy

$$t_{1,n}(0,0) = 1 - \theta_n r_{n+1}$$
$$t_{1,n}(0,1) = \theta_n r_{n+1}$$
$$t_{1,n}(1,0) = \theta_n (1 - r_{n+1}) \quad (7)$$
$$t_{1,n}(1,1) = 1 - \theta_n (1 - r_{n+1})$$

(Boehnke et al. 1991). Using the equations in expression 7 for the haploid case, we can write the more general equation for the c-ploid case according to Lange et al. (1995):

$$t_{c,n}(j,k) = \sum_{l=\max\{0,i+j-c\}}^{\min\{j,k\}} \left[\binom{j}{l} t_{1,n}(1,1)^l t_{1,n}(1,0)^{j-l} \times \binom{c-j}{k-l} t_{1,n}(0,1)^{k-l} t_{1,n}(0,0)^{c-j-k+l} \right]. \quad (8)$$

The distribution of obligate breaks for a single hybrid typed for M markers can be obtained by summing over k and w:

$$P(b) = \sum_{k=0}^{c} \sum_{w=0}^{1} P_M(b,k,w).$$

The probability $P(b)$ can be multiplied by the number of hybrids to obtain the expected number of hybrids with b obligate breaks.

In the presence of missing data, the same general framework for computing the distribution can be used if we consider only the typed markers. If locus A_i is the first typed locus in the order $(A_1, A_2, \ldots, A_M)$, then marker A_i is retained if there are no breaks between markers A_1 and A_i

and the fragment is retained, or if there is a break somewhere between markers A_1 and A_i and the fragment containing A_i is retained. Hence, the probability that A_i is retained in a hybrid is

$$r_i^* = r_1 \prod_{n=1}^{i-1} (1 - \theta_n) + \sum_{n=1}^{i-1} \theta_n r_{n+1} \prod_{j=n+1}^{i-1} (1 - \theta_j). \qquad (9)$$

In the equation for the initial condition of the recurrence relation (expression 5), we replace r_1 with r_i^*. If the nth and $(n + 1)$th typed markers are A_u and A_v, then in (8) we replace the haploid transition probabilities $t_{1,n}(i,j)$ between A_n and A_{n+1} with the haploid transition probabilities between A_u and A_v, $t_{1,uv}(i,j)$. These are obtained by replacing r_{n+1} with r_v and θ_n with $1 - \Pi_{k=u}^{v-1}(1 - \theta_k)$ in expression 7. The resulting polyploid transition probabilities may then be substituted into the recurrence relation (expression 6).

DISCUSSION

Whole-genome diploid radiation hybrids are becoming an important and convenient tool for mapping markers throughout the human genome. If the problem of low retention for high radiation dose panels cannot be solved by modification of experimental protocol, combining two or more diploid panels to create polyploid clones with higher per-hybrid retention appears to be a useful alternative. Although creating more hybrids and pooling them together requires a substantial initial outlay of resources, the bulk of the work for panels used to map many markers throughout the genome will be in marker typing and scoring. If polyploid panels are constructed by pooling hybrids, then our results suggest as a rule of thumb that a combination of ploidy c and chromosome-specific retention probability r yielding a per-hybrid retention probability $r_0 = 1 - (1 - r)^c$ of 0.5 will be nearly optimal for ordering markers and estimating distances between them. Our simulations with chromosome-specific retention probability $r = 0.08$ show that the ability to identify the true order and the support for the true order increase most from diploid to tetraploid ($c = 4$) hybrids (Fig. 1). The smaller gains achieved for 6- or 8-ploid hybrids may not warrant the creation of additional diploid panels for pooling.

Our simulation results suggest that analyzing polyploid data under a haploid model does not diminish the chance of correctly ordering mark-

ers but may give less reliable parameter estimates. These findings are futher supported by our experience with the diploid chromosome 4 data described in the accompanying paper (Lange et al. 1995). With the double-typed data we obtained exactly the same set of best orders using haploid and diploid analyses. The difference between the diploid and haploid maximum $\log_{10}$ likelihood for the best order was 0.18, and the difference in maximum $\log_{10}$ likelihood beteween the two best orders was slightly greater for the haploid than for the diploid analysis (1.241 and 1.218, respectively). The estimated map length was 3.356 Rays for the haploid and 3.135 Rays for the diploid analysis. Because haploid calculations are considerably faster, an efficient analysis strategy is to analyze polyploid data as haploid initially, and then use the polyploid analysis on the set of candidate orders to confirm the best orders and to obtain more accurate intermarker distance estimates.

In our examination of typing error we considered error rates in a range consistent with the discordance rates that we have seen in practice (Lange et al. 1995). We assumed that the two typings for double-typed data are independent and that the errors at adjacent markers are independent. These independence assumptions are unlikely to hold in practice. For example, if a marker is present in the hybrid at a level well below the threshold level required for a positive test, then two false-negative typings are likely. Similarly, if in a hybrid a DNA fragment containing a particular marker fails to amplify properly, then nearby markers on the same fragment are also likely to fail to amplify properly. Although our assumptions about error are flawed, they do allow us to account for typing errors at least approximately. Our examination of the effects of typing error on map length estimation clearly shows the importance of excluding data for which typing is uncertain. Missing data represent an approximately linear loss of information. In contrast, typing error rates as low as 0.5% result in greater than linear losses of information about the breakage probability, strongly reduce the ability to identify the true order, and artificially inflate the estimates of intermarker distances and map length. Our simulations suggest that explicitly allowing for typing error in data analyses results in less biased parameter estimates. However, even when correctly accounted for, typing errors reduce the power to identify the true locus order. In our analyses of simulated data we used the

correct error rates, that is, the error rates used in simulating the data. In practice, the overall error rate for a marker can be estimated as approximately one-half the rate of discordance between two independent typings of the hybrids, or a number of reasonable error rates can be tried, and the results of the analyses compared. Alternatively, error rates may be estimated along with the other model parameters (Lange et al. 1995). However, more accurate models for error may need to be developed to make this approach practical.

Our results show that there is somewhat greater power to identify the true order using 100 double-typed hybrids than using 200 single-typed hybrids and an assumed error rate. In contrast, the larger single-typed panel provides stronger support for the true order if a non-zero error rate is assumed. Including the possibility of error is more important than setting the error rates correctly: Error rates misspecified by a factor of two can still control map inflation. Furthermore, as long as error rates are 1% or lower, the larger single-typed panel gives us a clear advantage in building framework maps: For the cases we considered, we were able to include >70% more markers while only slightly increasing the probability of an incorrect map. Because a single whole-genome radiation hybrid panel has the potential to allow us to order markers anywhere in the human genome, it may be worthwhile to spend the additional resources to create one large panel of 200 or more hybrids and single type the hybrids for markers of interest. Rough estimates of error rates for various parts of the genome could be obtained by double typing a small set of markers on each chromosome, and these estimates could be used as guidelines when specifying error rates for analyses. We emphasize that a strategy of single typing larger sets of hybrids must be tested on real data before it is implemented on a broad scale, as our results based on simulation are based on a model for typing error that is only approximate. We also stress that single typing of 100 hybrids is not a viable strategy, because of its much poorer ordering accuracy (Table 3). Map building continues to be a daunting task. A tremendous amount of work is required both to create hybrids and to type and score them, as well as to verify the typings by retyping and rescoring. Further work needs to be done to investigate ways in which we can increase efficiency in map building without decreasing the quality of the maps we construct.

The computation of the distribution of obligate breaks is a convenient method for identifying "outlier" hybrids with many more obligate breaks than expected. Hybrids appearing to have many more obligate breaks than expected given no typing error should be retyped and rescored, either completely or at least at the loci delimiting the obligate breaks. In practice, we have found that in typical RH panels, we consistently observe more hybrids than expected (1) with zero breaks and (2) with many breaks. This apparent lack of fit of the observed number of obligate breaks to the distribution suggests that a hybrid-specific model for retention may be helpful; we are currently examining this issue further.

For simplicity, we assumed an equal fragment retention rate for all markers in our simulations and analyses. Our conclusions should apply similarly to data showing evidence of non-equal fragment retention as long as an appropriate fragment retention model is used in the analysis.

In summary, whole-genome radiation hybrids are a powerful tool for mapping and estimating distances between markers on any human chromosome. We hope that our conclusions will assist geneticists in understanding experimental design questions involving ploidy, retention rates, and error detection and correction. The methods for polyploid RH mapping discussed here are implemented in the FORTRAN 77 program package *RHMAP version 2.01* (Boehnke et al. 1994), available free of charge from Michael Boehnke.

ACKNOWLEDGMENTS

REFERENCES

Abel, K.J., M. Boehnke, M. Prahalad, P. Ho, W.L. Flejter, M. Watkins, J. Vanderstoep, S.C. Chandrasekharappa, F.S. Collins, T.W. Glover, and B.L. Weber. 1993. A radiation hybrid map of the BRCA1 region of chromosome 17q12-q21. *Genomics* **17:** 632–41.

Boehnke, M., K. Lange, and D.R. Cox. 1991. Statistical

methods for multipoint radiation hybrid mapping. *Am. J. Hum. Genet.* **49:** 1174–1188.

Boehnke, M., E. Hauser, K. Lange, K.L. Lunetta, J. Uro, and J. Vanderstoep. 1994. "RHMAP version 2.01: Statistical package for multipoint radiation hybrid mapping," unpublished computer program documentation.

Burmeister, M., S. Kim, E.R. Price, T. deLange, U. Tantravahi, R.M. Myers, and D.R. Cox. 1991. A map of the distal region of the long arm of human chromosome 21 constructed by radiation hybrid mapping and pulsed-field gel electrophoresis. *Genomics* **9:** 19–30.

Chakravarti, A. and J.E. Reefer. 1992. A theory for radiation hybrid (Goss-Harris) mapping: Application to proximal 21q markers. *Cytogenet. Cell. Genet.* **59:** 99–101.

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.

Falk, C.T. 1991. A simple method for ordering loci using data from radiation hybrids. *Genomics* **9:** 120–123.

Frazer, K.A., M. Boehnke, M.L. Budarf, R.K. Wolff, B.S. Emanuel, R.M. Myers, and D.R. Cox. 1992. A radiation hybrid map of the region on human chromosome 22 containing the neurofibromatosis type 2 locus. *Genomics* **14:** 574–584.

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.

———. 1977a. Gene transfer by means of cell fusion. I. Statistical mapping of the human X chromosome by analysis of radiation-induced gene segregation. *J. Cell. Sci.* **25:** 17–37.

———. 1977b. 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.

Haldane, J.B.S. 1919. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J. Genet.* **8:** 299–309.

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

Lange, K., M. Boehnke, D.R. Cox, and K.L. Lunetta. 1995. Statistical methods for polyploid radiation hybrid mapping. *Genome Res.* (this issue).

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.

Oshima, J., C.-E. Yu, M. Boehnke, J. Weber, S. Edelhoff, M.J. Wagner, D.E. Wells, S. Wood, C.M. Disteche, G.M. Martin, and G.D. Schellenberg. 1994. Integrated mapping analysis of the Werner syndrome region of chromosome 8. *Genomics* **23:** 100–113.

Richard, C.W., D.A. Withers, T.C. Meeker, S. Maurer, G.A. Evans, R.M. Myers, and D.R. Cox. 1991. A radiation hybrid map of the proximal long arm of human chromosome 11 containing the multiple endocrine neoplasia type 1 (MEN-1) and *bcl*-1 disease loci. *Am. J. Hum. Genet.* **49:** 1189–1196.

Richard, C.W., D.R. Cox, L. Kapp, J. Murnane, F. Cornelis, C. Julier, G.M. Lathrop, and M.R. James. 1993. A radiation hybrid map of human chromosome 11q22-q23 containing the ataxia-telangiectasia disease locus. *Genomics* **17:** 1–5.

Walter, M.A., D.J. Spillett, P. Thomas, J. Weissenbach, and P.N. Goodfellow. 1994. A method for constructing radiation hybrid maps of whole genomes. *Nature Genet.* **7:** 22–28.

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

Wilson, S.R. 1992. A statistical and numerical view of the chromosome 21 physical map data. *Cytogenet. Cell Genet.* **59:** 110–111.

Experimental design and error detection for polyploid radiation hybrid mapping.

K L Lunetta, M Boehnke, K Lange, et al.

References	This article cites 20 articles, 1 of which can be accessed free at: http://genome.cshlp.org/content/5/2/151.full.html#ref-list-1
License	
Email Alerting Service	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions