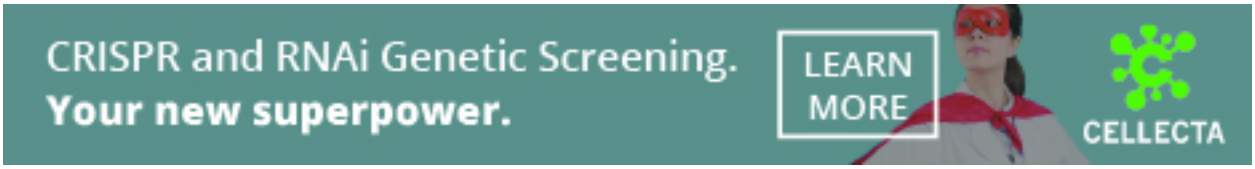
Statistical methods for gene map construction by fluorescence in situ hybridization.

S W Guo and W L Flejter

Genome Res. 1996 6: 1133-1148
Access the most recent version at doi:10.1101/gr.6.12.1133



RESEARCH

Statistical Methods for Gene Map Construction by Fluorescence in Situ Hybridization

Sun-Wei Guo[1,3] and Wendy L. Flejter[2]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109-2029; [2]Department of Pediatrics, University of Utah, Salt Lake City, Utah 84132

Fluorescence in situ hybridization (FISH) provides an efficient and powerful technique for ordering loci both on metaphase chromosomes and in less condensed interphase chromatin. Two-color metaphase FISH can be used to order pairs of loci relative to the centromere; two- and three-color interphase FISH can be used to accurately order trios of loci spaced within 1 Mb relative to one another. Loci separated by a distance >1–2 Mb exhibit chromatin loops that often give rise to a statistically significant but incorrect order. We derive Bayesian methods for selecting the best locus order based on microscopic evaluation for each of these types of FISH mapping data. We then describe how the results from several two- and three-locus analyses can be combined to evaluate the approximate posterior probability of a given multilocus order within the limits of the technology utilized. These methods directly address the question of interest: What is the probability that the inferred two-, three-, or multilocus order actually is correct? We illustrate our analysis methods by applying them to previously described FISH mapping data of 14 markers in the BRCA1 region on chromosome 17q12–q21. We also propose design strategies to order a group of closely spaced (<1 Mb) loci, two and three loci at a time, using a bisection strategy for two-color FISH data and a trisection strategy for three-color FISH data. These strategies have the best worst-case performance for ordering a new locus relative to a group of ordered loci and are nearly optimal for ordering a group of loci of unknown order. These, in conjunction with physical mapping strategies, provide efficient and reliable methods for gene map construction by FISH.

Identification of disease genes typically involves multiple steps. First, the gene is localized, by linkage analysis, to a specific chromosomal region, often flanked by two markers. Second, the region is saturated with additional markers to confirm and more closely localize recombinations and to develop a well-ordered physical map with overlapping contigs of YAC and cosmid clones. Finally, a number of complementary methods may be used to positionally clone the mutant gene of interest (Collins 1992).

One challenging aspect leading to the construction of a contig map is the ordering of multiple overlapping clones. The availability of sequence-tagged sites (STSs) (Olson et al. 1989) has made this procedure easier, providing a means to develop physical maps of entire chromosomes (Chumakov et al. 1992; Foote et al. 1992; Hudson et al. 1995). However, other methods for ordering a limited number of gene sequences and probes have also been developed. Radiation hybrid mapping (Cox et al. 1990; Boehnke et al. 1991) is one approach to order closely linked probes. Fluorescence in situ hybridization (FISH) has also emerged as an important approach for this purpose (Trask 1991b; Flejter et al. 1993; Wilke et al. 1994), although it has many other applications as well.

FISH involves the formation of a heteroduplex between DNA probes and chromatin targets on a microscope slide; the probes are visualized with fluorescent reporter molecules (Brandriff et al. 1991; Trask 1991a). Given the availability of fluorescent probe-labeling systems and detection reagents, FISH provides a means for the rapid localization of DNA segments to a specific chromosomal region in the absence of a long-range physical map.

The positions of single-copy DNA sequences, cloned in the form of YACs or cosmids, can be ordered efficiently with respect to metaphase

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

6:1133–1148 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

chromosomes with a resolution of >1 Mb (Trask et al. 1991a). To order sequences at a higher level of resolution (0.5–1 Mb), probes can be hybridized in situ to somatic or pronuclear interphase nuclei, where the chromatin is further stretched (Lawrence et al. 1990; Brandriff et al. 1991; Trask 1991a). As a result, markers can be ordered at a level of resolution intermediate between other mapping techniques such as genetic linkage analysis and restriction mapping. The use of FISH in gene mapping has increased rapidly in recent years (Lichter et al. 1990; Brandriff et al. 1991, 1992; Lebo et al. 1991; Barnes et al. 1992; Trask et al. 1992; Flejter et al. 1993; Wilke et al. 1994).

Although a variety of different experimental approaches can be used for gene mapping by FISH, in this paper we concentrate primarily on two situations: metaphase mapping of two probes relative to the centromere, where each probe is labeled by a different fluorochrome, and interphase mapping of three probes, where, again, each is labeled by a different fluorochrome. We also consider interphase mapping of three probes with two colors, two loci labeled with one color and one with another. In what follows, we refer to these as the two-color, three-color, and two-color–three-locus problems, respectively. In the Discussion, we describe modifications to allow for the analysis of multicolor FISH data.

For any mapping experiment, random noise (for metaphase mapping) or uncoiling and entangling or looping of the chromatin (for interphase mapping) can result in an apparent locus order different from the true locus order. Previous studies have shown that interphase mapping distance is most useful for estimating genomic separations <1 Mb (Trask 1991b; Trask et al. 1991a; Yokota et al. 1995). Therefore, a priori information regarding probe distance by restriction mapping, pulse-field electrophoreses, or some other mapping strategy should be considered in probe selection. In the absence of such information, probe distance can, in some cases, be estimated by metaphase FISH. For example, combinations of probes that cannot be resolved on standard or high-resolution chromosome preparations are likely to be ordered by interphase FISH with relatively high accuracy. Because replicate scorings are made and all possible orders are compatible with the observations, statistical inference of the locus order is necessary. Intuitively, within the limits of technology, the most likely locus order is the one that occurs most frequently in a set of metaphase or interphase cells scored. Given this, what we truly require is a method to evaluate the probability that the most frequent order is the correct order, given the data. For analysis of two-color mapping data, Trask et al. (1991a) proposed a large sample Z-test. Unfortunately, that test cannot be generalized to analyze data for three-color experiments. Moreover, the Z-test is a *fixed-sample* significance test, which is not valid if data are accrued in a sequential fashion. At best, the Z-test provides an answer to the question: Upon repeated sampling, how often would the two-locus order inferred in this way be correct? The Z-test cannot answer the more relevant question: What is the probability that the inferred two-, three-, or multilocus order actually is correct?

In this paper we derive statistical methods to answer this question. First, we derive a statistical method to evaluate the uncertainty in selecting the best order for the two-color, three-color, and two-color–three-locus problems. Second, we derive a method to evaluate the probability for a given multilocus order constructed through a sequence of two-color, three-color, and two-color–three-locus FISH experiments using sets of probes that lie within the limits of resolution for metaphase and interphase FISH mapping. Finally, in the realm of experimental design, we propose bisection and trisection strategies for ordering sets of loci using the minimum number of two-color and three-color FISH mapping experiments.

We have written a set of computer programs in C called FISHMAP to carry out the analyses described in this paper. FISHMAP is available free of charge from the first author, either through e-mail from swguo@sph.umich.edu or upon receipt of a 3½-inch diskette formatted for an IBM PC or compatible, along with a self-addressed mailer.

RESULTS

We applied the proposed methods to two- and three-color mapping data on 15 markers on chromosome 17q (Flejter et al. 1993). Twelve two-color and 12 three-color FISH experiments were performed. Rather than using a bi- or trisection strategy, an ad hoc strategy was used in choosing the loci for each experiment. The data are summarized in Table 1 for two-color metaphase mapping data and in Table 2 for three-color interphase mapping data. Because in no experiment was the order of PPY and p131 resolved, these

Table 1. Posterior Probabilities and Z-test Results for Two-locus Metaphase FISH Data

No.	Data		$P(R\|\mathbf{n})$	Z	P	Best order[a]
1	22	2	0.999990	4.08	.00005	TOP2-RNU2
2	18	2	0.999889	3.58	.00035	RNU2-PPY/p131
3	34	0	1.000000	5.83	.00000	GAS-PPY/p131
4	14	0	0.999969	3.74	.00018	PPY/p131-WNT3
5	6	4	0.725586	0.63	.52709	PPY/p131-EPB3
6	20	6	0.997038	2.75	.00604	17HSD-EPB3
7	20	1	0.999995	4.15	.00003	17HSD-PPY/p131
8	20	0	1.000000	4.47	.00001	17HSD-MFD188
9	32	0	1.000000	5.66	.00000	PPY/p131-MFD188
10	18	2	0.999889	3.58	.00035	EPB3-MFD188
11	20	2	0.999967	3.84	.00012	MFD188-WNT3
12	33	3	1.000000	5.28	.00000	MFD188-GP3A

From Flejter et al. (1993).
[a]The best locus order is arranged from left to right, proximal to distal.

two loci will be regarded in what follows as a single locus, designated as PPY/p131.

Table 1 presents posterior probabilities and, for comparison purposes, Z-test results for metaphase data, along with the best orders. Although high posterior probabilities are in general associated with small P values, the interpretations of the two procedures are completely different. For example, given data $\mathbf{n} = (18,2)$, $P(R_1|\mathbf{n}) = 0.999889$, $Z = 3.578$, $P = 0.000347$. $P(R_1|\mathbf{n})$ is the posterior probability that the selected order is correct. The P value, however, is the probability that one observes $n_2 = 2$ or *more extreme* cases: (19,1) and (20,0) or (2,18), (1,19), and (0,20).

Table 2 shows the posterior probabilities and the best locus orders for the three-color interphase data. On the basis of these results and previously published data (Fain 1992) that suggest the order cen–WNT3–HOX2–tel, Flejter et al. (1993) concluded that the best locus order for the 15 loci is: cen–THRA1–TOP2–GAS–OF2–17HSD–248yg9–RNU2–OF3–PPY/p131–EPB3–MFD188–WNT3–HOX2–GP3A–tel, which we represent numerically in Figure 1 as 1–2–3–4–5–6–7–8–9–10–

Table 2. Posterior Probabilities Three-locus Interphase FISH Data

No.	Data			$P_E(R\|\mathbf{n})$	$P_G(R\|\mathbf{n})$	Best order
1	15	3	2	0.999719	0.997164	GAS-TOP2-THRA1
2	19	2	1	0.999999	0.999929	TOP2-GAS-17HSD
3	14	2	0	0.999982	0.998796	17HSD-Of2-GAS
4	20	8	6	0.995919	0.985402	RNU2-OF2-GAS
5	14	3	1	0.999758	0.995993	GAS-17HSD-RNU2
6	18	2	2	0.999993	0.999780	RNU2-248yg9-17HSD
7	23	2	0	1.000000	0.999995	PPY/p131-RNU2-17HSD
8	27	3	0	1.000000	0.999998	PPY/p131-OF3-RNU2
9	17	2	1	0.999996	0.999761	OF3-PPY/p131-EPB3
10	9	3	1	0.984101	0.949588	EPB3-PPY/p131-RNU2
11	12	7	6	0.829697	0.816252	PPY/p131-EPB3-MFD188
12	16	3	2	0.999879	0.998361	WNT3-HOX2-GP3A

From Flejter et al. (1993). $P_E(\mathbf{R}|\mathbf{n})$ and $P_G(\mathbf{R}|\mathbf{n})$ are posterior probabilities calculated for the equal and general error probability models, respectively.

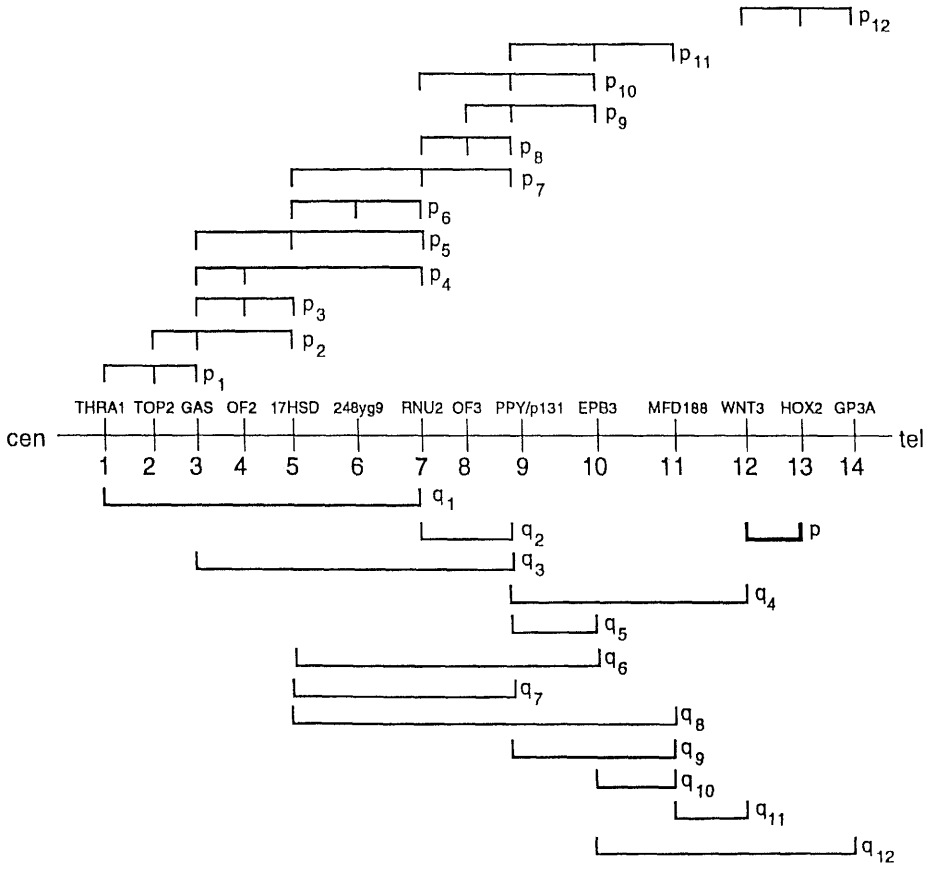


Figure 1 The locus order determined by 24 FISH experiments and one external result (Flejter et al. 1993). The q_i's are posterior probabilities from two-color metaphase FISH experiments (Table 1), and the p_i's are posterior probabilities from three-color interphase FISH experiments (Table 2). *P* denotes a result from an external data source (Fain 1992). Numerical representations of the 14 loci are given under the locus names.

11–12–13–14. Figure 1 depicts the locus order, along with the posterior probabilities conditional on the two- and three-color FISH data. Note that if the order cen–WNT3–HOX2-tel had not been determined previously, there would be two equally plausible orders: the one specified previously and the other identical to it but with WNT3–HOX2–GP3A inverted.

To allow an explicit comparison of complete enumeration and simulated annealing, we first used both methods for 12 loci, discarding loci HOX2(13) and GP3A(14), and two-color experiment 12 and three-color experiment 12. Under the general error probability model, complete enumeration and summation over all locus orders yields a posterior probability of 0.972 for the best order, whereas simulated annealing and summation over the 200 best identified orders gives 0.976 (Table 3). The upper and lower panels of Table 3 give the 20 best locus orders by complete enumeration and simulated annealing, respectively. The numbers in the odds ratio column give the ratios of the probabilities of the best locus order and other nearly best locus orders. It can be seen that, although the 20 best locus orders are different, simulated annealing was able to identify the 13 best locus orders and resulted in similar posterior probabilities for the same orders. Complete enumeration and simulated annealing required ~31 hr and 52 min on a SUN SPARC2 workstation, respectively. The equal error probability model gave similar results (data not shown).

Table 4 lists the 10 best orders identified by one run of simulated annealing using all 24 FISH experiment results but not the published result (Fain 1992). Although the 10 best locus orders are slightly different under the equal and general error probability models, both models demonstrate that there are two equally best orders, as expected. Each of these analyses required ~50 min on our SUN SPARC 2.

Table 3. Twenty Best Locus Orders for 12 Loci Under the General Error Probability Model

No.	Order												$P(R\|n)$	Odds ratio
1	1	2	3	4	5	6	7	8	9	10	11	12	.9723	1
2	2	1	3	4	5	6	7	8	9	10	11	12	.0021	454
3	2	3	4	5	6	7	8	9	10	11	1	12	.0021	454
4	2	3	1	4	5	6	7	8	9	10	11	12	.0021	454
5	2	3	4	5	6	7	1	8	9	10	11	12	.0021	454
6	2	3	4	5	6	1	7	8	9	10	11	12	.0021	454
7	2	3	4	5	1	6	7	8	9	10	11	12	.0021	454
8	2	3	4	5	6	7	8	1	9	10	11	12	.0021	454
9	2	3	4	5	6	7	8	9	1	10	11	12	.0021	454
10	2	3	4	1	5	6	7	8	9	10	11	12	.0021	454
11	2	3	4	5	6	7	8	9	10	1	11	12	.0021	454
12	2	3	4	5	6	7	8	9	10	11	12	1	.0021	454
13	1	2	3	5	6	4	7	8	9	10	11	12	.0011	851
14	1	2	3	5	4	6	7	8	9	10	11	12	.0011	851
15	1	2	3	4	6	5	7	8	9	10	11	12	.0001	9071
16	1	2	3	4	5	7	6	8	9	10	11	12	.0001	9071
17	1	2	3	6	5	4	7	8	9	10	11	12	.0001	9071
18	1	6	2	3	4	5	7	8	9	10	11	12	.0001	9071
19	6	1	2	3	4	5	7	8	9	10	11	12	.0001	9071
20	1	2	3	4	5	7	8	6	9	10	11	12	.0001	9071
1	1	2	3	4	5	6	7	8	9	10	11	12	.9761	1
2	2	3	4	5	6	7	8	9	10	11	12	1	.0022	454
3	2	3	4	5	6	7	8	9	10	11	1	12	.0022	454
4	2	3	4	5	6	7	8	9	10	1	11	12	.0022	454
5	2	1	3	4	5	6	7	8	9	10	11	12	.0022	454
6	2	3	4	5	6	7	8	9	1	10	11	12	.0022	454
7	2	3	4	5	6	7	1	8	9	10	11	12	.0022	454
8	2	3	4	5	1	6	7	8	9	10	11	12	.0022	454
9	2	3	4	1	5	6	7	8	9	10	11	12	.0022	454
10	2	3	1	4	5	6	7	8	9	10	11	12	.0022	454
11	2	3	4	5	6	7	8	1	9	10	11	12	.0022	454
12	2	3	4	5	6	1	7	8	9	10	11	12	.0022	454
13	3	4	5	6	2	7	8	9	10	11	12	1	.0000	95675
14	3	4	5	6	2	7	8	9	10	11	1	12	.0000	95675
15	3	4	5	2	6	7	8	9	10	11	12	1	.0000	95675
16	3	4	2	5	6	7	8	9	10	11	12	1	.0000	95675
17	3	4	5	2	6	7	8	9	10	1	11	12	.0000	95675
18	3	4	2	5	6	7	8	9	10	1	11	12	.0000	95675
19	3	4	2	5	6	7	8	9	10	11	1	12	.0000	95675
20	3	2	4	5	6	7	8	9	10	11	1	12	.0000	95675

The upper and lower panels list results from complete enumeration and simulated annealing, respectively. $K = 200$ is used for simulated annealing to estimate the posterior probabilities. The odds ratio compares the probability of the best locus order with that for other nearly best locus orders.

To mimic a highly reliable result from an external source that essentially forces the order cen–WNT3–HOX2 (cen–12–13), we assigned a two-color FISH result of (34, 0) for loci 12 and 13; this translates to a posterior probability of 1.000000 for the order cen–12–13. We then used simulated annealing again to identify the best locus orders. The results (Table 5) show that, when the external information is incorporated, both models suggest that the posterior probability of the inferred locus order is substantially >0.99.

Table 4. Ten Best Locus Orders for 14 Loci Identified by Simulated Annealing

No.	Order														$P(R\|\mathbf{n})$	Odds ratio
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	.49774	1
2	1	2	3	4	5	6	7	8	9	10	11	14	13	12	.49774	1
3	2	3	1	4	5	6	7	8	9	10	11	14	13	12	.00110	454
4	2	1	3	4	5	6	7	8	9	10	11	12	13	14	.00110	454
5	2	3	4	1	5	6	7	8	9	10	11	12	13	14	.00110	454
6	2	3	1	4	5	6	7	8	9	10	11	12	13	14	.00110	454
7	3	4	5	6	2	1	7	8	9	10	11	14	13	12	.00001	95675
8	3	4	5	2	6	1	7	8	9	10	11	14	13	12	.00001	95675
9	3	4	5	6	2	7	1	8	9	10	11	12	13	14	.00001	95675
10	3	4	2	5	6	1	7	8	9	10	11	14	13	12	.00001	95675
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	.49968	1
2	1	2	3	4	5	6	7	8	9	10	11	14	13	12	.49968	1
3	1	2	3	13	4	5	6	7	8	9	10	11	14	12	.00003	18136
4	1	13	2	3	4	5	6	7	8	9	10	11	14	12	.00003	18136
5	1	2	3	4	5	6	7	13	8	9	10	11	14	12	.00003	18136
6	1	2	3	4	13	5	6	7	8	9	10	11	12	14	.00003	18136
7	1	2	3	4	5	6	7	8	9	10	13	11	14	12	.00003	18136
8	1	2	3	4	5	6	13	7	8	9	10	11	12	14	.00003	18136
9	1	2	3	4	5	6	7	8	9	13	10	11	12	14	.00003	18136
10	1	2	3	4	5	13	6	7	8	9	10	11	14	12	.00003	18136

The upper and lower panels list results under the general error probability model and equal error probability model, respectively. $K = 200$ is used in all computations.

DISCUSSION

We have presented Bayesian statistical methods for selecting the best locus order for two-color and three-color FISH mapping experiments and for evaluating the posterior probability of a multilocus map constructed using FISH data. These methods have several important advantages. First and most important, for individual mapping experiments or for complete maps in which the number of loci to order is not too large, these methods directly address the question of interest: What is the probability that a particular locus order is correct? This answer is more readily interpretable than a P value, which relies on the ideas of more extreme results and repeated sampling; for example, for two-color data, the posterior probability of choosing one particular order without data [i.e., $\mathbf{n} = (0,0)$] $P(R|\mathbf{n}) = ½$, which agrees with common sense. (Note that Z is undefined in this case.) For $\mathbf{n} = (m,m)$, where m is any positive integer, $P(R|\mathbf{n}) = ½$ again agrees with common sense. (Note that $Z = 0$, $P = 1$.) Similar cases can be also observed for three-color data. Even for very large numbers of loci, an evaluation of the odds ratio between the best and next-best locus orders can be obtained straightforwardly.

Second, these methods can answer the locus-orientation question solely on the basis of the evidence at hand or by incorporating prior information regardless of how the experiments are conducted, as demonstrated by our chromosome 17 example. Third, the analysis for individual experiments are computationally straightforward, although combining experimental results to obtain evidence for an entire map can be time consuming if the number of loci is large.

We also have presented a bisection strategy for map construction by two-color metaphase experiments and a trisection strategy for map construction by three-color FISH experiments. For placing a new locus in an existing map, both have the best worst-case performance in their application range. The sequential versions of the bi- and trisection strategies are nearly optimal when used to build a map from a set of unordered loci.

Note that modest redundancy in nonoptimal designs is sometimes desirable. We point out that the proposed design strategy conflicts in no way with the redundancy principle. The bi- and tri-

Table 5. Ten Best Locus Orders for 14 Loci Identified by Simulated Annealing Using 24 FISH Experiments and One External Result

No.	Order														$P(R\|n)$	Odds ratio
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	.99839	1
2	1	2	3	5	4	6	7	8	9	10	11	12	13	14	.00117	851
3	6	1	2	3	4	5	7	8	9	10	11	12	13	14	.00011	9071
4	1	6	2	3	4	5	7	8	9	10	11	12	13	14	.00011	9071
5	1	2	6	3	4	5	7	8	9	10	11	12	13	14	.00011	9071
6	1	2	3	6	4	5	7	8	9	10	11	12	13	14	.00011	9071
7	2	6	1	3	4	5	7	8	9	10	11	12	13	14	.00000	4.1×10^6
8	2	1	6	3	4	5	7	8	9	10	11	12	13	14	.00000	4.1×10^6
9	2	6	3	1	4	5	7	8	9	10	11	12	13	14	.00000	4.1×10^6
10	2	6	3	1	5	4	7	8	9	10	11	12	13	14	.00000	3.5×10^9
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	.99801	1
2	2	3	4	5	6	7	8	9	10	11	1	12	13	14	.00015	6524
3	2	3	4	5	6	7	8	9	10	1	11	12	13	14	.00015	6524
4	2	3	4	5	6	7	8	9	1	10	11	12	13	14	.00015	6524
5	2	3	4	5	6	7	8	1	9	10	11	12	13	14	.00015	6524
6	2	3	4	5	6	7	8	9	10	11	12	1	13	14	.00015	6524
7	2	3	4	5	6	7	8	9	10	11	12	13	1	14	.00015	6524
8	2	3	4	5	6	7	8	9	10	11	12	13	14	1	.00015	6524
9	2	3	4	5	6	7	1	8	9	10	11	12	13	14	.00015	6524
10	2	3	4	5	6	1	7	8	9	10	11	12	13	14	.00015	6524

The upper and lower panels list results under the general error probability model and equal error probability model, respectively. $K = 200$ is used in all computations.

section strategies have the best performance *in the worst-case scenario*, which means that, in most cases, the strategy is somewhat redundant. We also point out that the noninformative prior assumption provides some redundancy, as it yields slightly conservative results.

For three-color mapping experiments, there are three possible locus orders, two of which are incorrect. We have considered two models in this situation: One that assumes the probability of incorrectly observing these orders is the same and the second that allows these probabilities to differ. Because the distances between loci generally will be different, we would expect the corresponding error probabilities also to be different. This makes the general error probability model more reasonable, and we regard it as the model of choice.

Given two- and three-color mapping, we might consider more general k-color mapping ($k > 3$). Multiple colors now are possible, and the statistical methods we have described easily can be generalized to any number of colors k. However, both experimental and statistical considerations suggest that going much beyond three loci is unlikely to be profitable. First, as the number of loci increases, so too do the number of unscorable observations. Even for three-color mapping experiments, typically 20–30% of observations cannot be unequivocally scored. Second, as the number of loci k increases, the number of locus orders $k!$ or $k!/2$ increases much more rapidly, so that substantially larger numbers of scorable observations are required to infer a locus order. Third, larger numbers of loci result in an increase in the cost of each experiment.

Throughout the paper we have assumed that the true locus order is the one observed most frequently. In other words, random errors in scoring have been assumed. This may not be always true, as observed by Yokota et al. (1995), who reported that nonrandom loops and folds in interphase chromatin give rise to a most frequently observed but wrong order. Obviously, our method will not be applicable to this situation. The correspondence between a statistically significant interphase order and genomic order depends on the nature of DNA folding within the interphase

nucleus. However, for loci spaced within 1 Mb, our assumption seems to be valid (Trask 1991b; Trask et al. 1991a; Yokota et al. 1995).

We also have assumed a noninformative prior for the error rates in the probability calculations. Although the noninformative prior is unlikely to be true in reality and, in principle, one can, at the cost of numerical complexity, use some informative prior and incorporate it in the probability calculation, there is no guarantee that the informative prior used will be true for all experiments and, as a consequence, will be accepted by all investigators. Furthermore, use of informative priors will make even approximate evaluation of the posterior probability of a map intractable as equation 11 holds only when a noninformative prior is assumed.

Because in most cases the error rates are quite small, the use of a noninformative prior tends to be conservative, which may be desirable in practice. The use of noninformative prior is consistent with the philosophy of "not putting all one's eggs into a single basket." There are four additional points that support using a noninformative prior. First, a locus order, once established with high confidence and confirmed by other independent groups, will become a consensus order and is unlikely to be investigated further, unless there is strong evidence against it. As a result, the data at hand often will not be sufficient to construct a prior distribution. Second, use of a noninformative prior will protect against inconsistencies owing to use of different priors, as the error rate is likely to be different for different labs. Third, use of noninformative prior requires minimum human intervention and greatly simplifies computation, which is crucial for a procedure to be followed strictly in practice. A complicated statistical procedure, however elegant and correct, is prone to human errors and is less likely to be used. Fourth, use of a noninformative prior is actually consistent with the conditional frequentists' philosophy that makes inference conditional on data (Berger 1985; Berry 1987).

Van den Engh et al. (1992) proposed to order loci and estimate physical distances between them from two-locus FISH mapping data in the context of a random walk polymer-chain model. Their idea is simple yet innovative. Swollen and immersed in a solution, the chromatin that harbors two loci, *A* and *B*, *n* bp away, resembles very much a polymer chain. Furthermore, immobilized and measured on a microscope slide, the distribution of the the physical distance between *A* and *B* is the same as that of the distance between two end points of the chain projected onto a *random* plane. This problem has been well studied (e.g., see Flory 1989). As long as one knows the projected unit length on the microscope slide for a single base pair, which can be, in principle, measured through experiments, the length of the physical distance between *A* and *B* can be estimated from two-locus FISH data. By analogy in the three-locus case, if one knows the joint distribution of three sides of the triangle, projected from a three-dimension space onto a *random* plane, with vertices *A*, *B*,, and *C*, one also can estimate the genomic distances between loci *A*, *B*, and *C* and hence, infer the order of the loci. Unfortunately, finding this distribution is not a trivial problem.

Intuitively, the probe–probe distance information would be helpful to infer locus orders. In some experiments, such distance information might be available. Our proposed method retains only the locus order information. This is, of course, owing partly to our desire for mathematical tractability, and it may not be efficient. However, because the relationship between distance and locus order has only been described empirically and is far from sufficient for modeling purposes, incorporation of such information, if available, into our method is extremely difficult, if not impossible. Thus, taking the distance data into account may have to await for more thorough investigation of such a relationship.

In summary, we have developed Bayesian statistical methods for the analysis of two- and three-locus FISH mapping data and a method to combine such results to calculate the posterior probability of a multilocus order. We recommend ordering genes by a bi- or trisection strategy, for two-color or three-color mapping experiments, respectively. These strategies, in conjunction with statistical methods proposed in this paper, provide simple, efficient, and reliable methods for gene map construction by FISH.

METHODS

Two-color Metaphase FISH Mapping

Suppose two loci, *A* and *B*, are to be ordered on metaphase chromosomes with respect to a specific reference point, usually the centromere. Using FISH, each probe is labeled with a different fluorochrome. Owing to condensation of the chromosome and the distance between the two probes, the position of *A* relative to *B* can be scored as proximal, distal, or even (that is, one on top of another or

side by side). Let n_1, n_2, and n_0 be the numbers of such observations, respectively. "Even" observations do not provide information on the order of *A* and *B* and are discarded.

Three-color Interphase FISH Mapping

For a DNA segment containing three loci, *A*, *B*, and *C*, there are total of 3 = 6 possible orders. Because in interphase mapping there is no visible orientation along the chromatin fiber, reversed orders such as *ABC* and *CBA* cannot be distinguished. Thus, there are only 3!/2 = 3 distinct orders. For notational convenience, we will denote these three orders as *ABC*, *BAC*, and *ACB*, with the understanding that, when orientation is considered, each represents two equally likely locus orders, one being the reverse of the other. Using three-color FISH, each locus is highlighted with one of three colors, such as red, green, and orange. Owing to randomness in chromatin structure, the apparent order may be *ABC*, *BAC*, *ACB*, or uncertain. Let n_1, n_2, n_3, and n_0 be the numbers of observations so scored. "Uncertain" observations are discarded.

Two-color–Three-locus Interphase FISH Mapping

In contrast to three-color FISH mapping, now the three loci are labeled with only two colors. Suppose *B* is labeled in green and *A* and *C* are labeled in red. For all scorable observations, there are two possible outcomes: either *B* is flanked by *A* and *C* or *B* is to one side of *A* and *C*. The former outcome implies the order *ABC* and the latter, denoted as *(A,C)B*, implies two equally probable orders, *ACB* and *BAC*. We will let n_1 and n_2 be the numbers of observations so scored.

Statistical Methods for Selecting the Best Locus Order

Because it is clear that we will select the locus order with the largest number of observations as the most likely, the real task is to quantify the degree of uncertainty in making this inference. Statistically, the two- and three-color ordering problems are special cases of a more general problem of selecting the most probable multinomial event, which has been studied extensively (e.g., Alam 1971; Ramey and Alam 1979). Methods to solve this problem generally assume that the probability ratio comparing the most probable outcome (correct locus order) with the next most probable outcome is greater than some known constant $\gamma > 1$. The most efficient of these methods are sequential (Alam 1971; Ramey and Alam 1979; Bechhofer and Goldsman 1985), assuming that observations are obtained one at a time. Unfortunately, one typically does not have definitive knowledge about probability ratios before an experiment. More fundamentally, because generating observations for map construction by FISH is substantially more expensive than scoring them, a strategy based on scoring one observation at a time is not practical.

Instead, we propose a Bayesian method. The method can be further extended to approximate the posterior probability of a locus order for a set of loci given a series of data sets (see below).

Suppose that each scoring results in an explicit observation of one of k orderings, where $k = 2$ for a two-color experiment, $k = 3$ for a three-color experiment, and all inconclusive observations are excluded. Suppose, furthermore, we observe data $\mathbf{n} = (n_1,...,n_k)$. Without loss of generality, suppose $n_1 = \max(n_1,...,n_k)$. Naturally, one would choose the corresponding order R_1 as most probable. Thus, given $\mathbf{n}$, we wish to evaluate the probability that the selected order R_1 is the correct locus order. In statistical terms, we want to evaluate $P(R_1|\mathbf{n})$. A priori, $P(R_i) = 1/k$, for $1 \leq i \leq k$. Hence, by Bayes's theorem,

$$P(R_1|\mathbf{n}) = \frac{P(\mathbf{n}|R_1)}{\Sigma_i P(\mathbf{n}|R_i)}, \tag{1}$$

and the problem of evaluating the probability of correct selection reduces to calculating the conditional probability $P(\mathbf{n}|R)$ of the data $\mathbf{n}$ for a given order R.

Method for the Two-color Analysis

Let θ be the probability of observing an incorrect order. If we assume the true locus order is the most likely to be observed, an assumption seems to hold well for loci spaced within 1 Mb, θ must satisfy $0 \leq \theta < \frac{1}{2}$. Given true order R_1 and θ, $\mathbf{n} = (n_1,n_2)$ is a sample from the binomial distribution

$$P(\mathbf{n}|R_1,\theta) = \frac{n!}{n_1!n_2!}(1-\theta)^{n_1}\theta^{n_2}$$

where $n = n_1 + n_2$.

Because θ is unknown and varies from experiment to experiment depending on experimental conditions and the distances between the two loci, we assume that θ is a random variable uniformly distributed between 0 and ½. This amounts to assigning a noninformative prior to θ. Therefore, because

$$P(\mathbf{n}|R_1) = \int_0^{1/2} P(\mathbf{n}|R_1,\theta)\, f(\theta)d\theta,$$

$$P(R_1|\mathbf{n}) = \frac{\int_0^{1/2}(1-\theta)^{n_1}\theta^{n_2}d\theta}{\int_0^{1/2}(1-\theta)^{n_1}\theta^{n_2}d\theta + \int_0^{1/2}(1-\theta)^{n_2}\theta^{n_1}d\theta} = \frac{J_{1/2}(n_2+1, n_1+1)}{B(n_2+1, n_1+1)} \tag{2}$$

Here $J_x(m,n) = \int_0^x t^{m-1}(1-t)^{n-1}dt$ is the incomplete beta function, and $B(m,n) = J_1(m,n)$ is the (complete) beta function.

Although $J_x(m,n)$ can be evaluated by expanding $(1-t)^{n-1}$ for $n > 1$, the results may be numerically unstable for large n owing to underflow. To avoid this problem, we use the recursive formula:

$$J_x(m,n) = \frac{1}{m}x^m(1-x)^{n-1} + \frac{n-1}{m}J_x(m+1, n-1)$$

until $n = 1$. To evaluate $J_x(m,n)$ for $m < n$, note that

$$J_x(m,n) = B(m,n) - J_{1-x}(n,m).$$

Method for the Three-color Analysis

Given three loci and three possible orders, there are now two error probabilities, θ_1 and θ_2. If we continue to assume the true locus order is the most likely to be observed, θ_1 and θ_2 must satisfy the constraint C: $0 \leq \theta_1$, $\theta_2 < 1 - \theta_1 - \theta_2$ or equivalently,

$$\text{C:} \quad 0 \leq \theta_2 < 1 - 2\theta_1, \quad 0 \leq \theta_2 < (1 - \theta_1)/2.$$

As for the two-color analysis, we assume that the error probabilities are uniformly distributed on their set of possible values in S_C, where S_C is the region satisfying constraint C. Thus, the joint density

$$f(\theta_1,\theta_2) = \begin{cases} 6 \text{ if } (\theta_1,\theta_2) \in S_C \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

This amounts to assigning a noninformative prior to θ_1 and θ_2.

Given order R_1 and (θ_1,θ_2), $\mathbf{n} = (n_1,n_2,n_3)$ is a sample from the trinomial distribution

$$P(\mathbf{n}|R_1,\theta_1,\theta_2) = \frac{n!}{n_1!n_2!n_3!}(1 - \theta_1 - \theta_2)^{n_1}\theta_1^{n_2}\theta_2^{n_3}$$

where $n = n_1 + n_2 + n_3$. Therefore,

$$P(\mathbf{n}|R_1) = 6\int_{(\theta_1,\theta_2)\in S_C} P(\mathbf{n}|R_1,\theta_1,\theta_2)d\theta_1 d\theta_2$$

and

$$P(R_1|\mathbf{n}) = \frac{\int_{(\theta_1,\theta_2)\in S_C} P(\mathbf{n}|R_1,\theta_1,\theta_2)d\theta_1 d\theta_2}{\sum_{i=1}^{3}\int_{(\theta_1,\theta_2)\in S_C} P(\mathbf{n}|R_i,\theta_1,\theta_2)d\theta_1 d\theta_2} \tag{4}$$

If we define

$$\begin{aligned}
I(n_1,n_2,n_3) &= \int_{(\theta_1,\theta_2)\in S_C}(1 - \theta_1 - \theta_2)^{n_1}\theta_2^{n_2}\theta_1^{n_3}\,d\theta_1\,d\theta_2 \\
&= \sum_{l=0}^{n_1}\frac{n_1!}{l!(n_1 - l)!}(-1)^l\int_{(\theta_1,\theta_2)\in S_C}(1 - \theta_1)^{n_1 - l}\theta_2^{n_2+l}\theta_1^{n_3}\,d\theta_1\,d\theta_2 \\
&= \sum_{l=0}^{n_1}\frac{n_1!}{l!(n_1 - l)!}(-1)^l\int_0^{1/3}\theta_2^{n_2}(1 - \theta_2)^{n_1 - l} \\
&\quad \left(\int_0^{1/2(1-\theta_2)}\theta_1^{n_3+l}\,d\theta_1\right)d\theta_2 \\
&\quad + \sum_{l=0}^{n_1}\frac{n_1!}{l!(n_1 - l)!}(-1)^l\int_0^{1/3}\theta_1^{n_3}(1 - \theta_1)^{n_1 - l} \\
&\quad \left(\int_{1/3}^{1/2(1-\theta_1)}\theta_2^{n_2+l}\,d\theta_2\right)d\theta_1 \\
&= \sum_{l=0}^{n_1}\frac{n_1!}{l!(n_1 - l)!}(-1)^l\left[\frac{(1/2)^{n_3+l+1}}{n_3 + l + 1}J_{1/3}(n_2 + 1, n - n_2 + 2)\right. \\
&\quad + \frac{(1/2)^{n_2+l+1}}{n_2 + l + 1}J_{1/3}(n_3 + 1, n - n_3 + 2) \\
&\quad \left. - \frac{(1/3)^{n_2+l+1}}{n_2 + l + 1}J_{1/3}(n_3 + 1, n_1 - l + 1)\right]
\end{aligned}$$

and similarly define $I(n_2, n_1, n_3)$ and $I(n_3, n_1, n_2)$, then

$$P(R_1|\mathbf{n}) = \frac{I(n_1,n_2,n_3)}{I(n_1,n_2,n_3) + I(n_2,n_1,n_3) + I(n_3,n_1,n_2)}. \tag{5}$$

Alternatively, we could assume $\theta_1 = \theta_2 = \theta/2$, where θ is uniformly distributed on the interval $0 \leq \theta < 2/3$. By an analogous argument,

$$\begin{aligned}
P(R_1|\mathbf{n}) &= \frac{\int_0^{2/3}(1 - \theta)^{n_1}(\theta/2)^{n - n_1}\,d\theta}{\sum_{i=1}^{3}\int_0^{2/3}(1 - \theta)^{n_i}(\theta/2)^{n-n_i}d\theta} \\
&= \frac{2^{n_1}J_{2/3}(n - n_1 + 1, n_1 + 1)}{\sum_i 2^{n_i}J_{2/3}(n - n_i + 1, n_i + 1)}
\end{aligned} \tag{6}$$

We refer to this latter model as the equal error probability model in contrast to the previous general error probability model.

Method for the Two-color–Three-locus Analysis

The general method for the three-color analysis can be readily extended to deal with a two-color–three-locus experiment. Consider again three loci, *A*, *B*, and *C*, and suppose a two-color–three-locus FISH experiment yields n_1 observations of *ABC* and n_2 observations of *(A,C)B*. Because $P(\mathbf{n}|BAC) = P(\mathbf{n}|ACB)$,

$$P(ABC|\mathbf{n}) = \frac{P(\mathbf{n}|ABC)}{P(\mathbf{n}|ABC) + 2P(\mathbf{n}|ACB)} \tag{7}$$

$$P(ACB|\mathbf{n}) = P(BAC|\mathbf{n}) = \frac{P(\mathbf{n}|ACB)}{P(\mathbf{n}|ABC) + 2P(\mathbf{n}|ACB)} \tag{8}$$

Although orders *ACB* and *BAC* cannot be distinguished in *(A,C)B* scorings because only two colors are being used, we can still let the theoretical probabilities of observing *ABC*, *ACB*, and *BAC* in one scoring be θ_1, θ_2, and θ_3, respectively, if three colors were being used. To calculate $P(\mathbf{n}|ABC)$, notice that, if *ABC* is the true order, θ_2 and θ_3 are error probabilities, which implies the constraint D on θ_2 and θ_3

$$\text{D:} \quad 0 \leq \theta_2 < (1 - \theta_3)/2 \quad 0 \leq \theta_3 < (1 - \theta_2)/2.$$

Again, we assign a noninformative prior on (θ_2,θ_3). Thus,

$$\begin{aligned}
P(\mathbf{n}|ABC) &= \int_{(\theta_2,\theta_3)\in S_D} P(\mathbf{n}|ABC,\theta_2,\theta_3)\,f(\theta_2,\theta_3)d\theta_2 d\theta_3 \\
&= 6\frac{n!}{n_1!n_2!}\int_{(\theta_2,\theta_3)\in S_D}(1 - \theta_2 - \theta_3)^{n_1}(\theta_2 + \theta_3)^{n_2}\,d\theta_2 d\theta_3
\end{aligned}$$

Using the change of variables $s = \theta_2 + \theta_3$, $t = \theta_2 - \theta_3$ gives

$$\begin{aligned}
P(\mathbf{n}|ABC) &= 3\frac{n!}{n_1!n_2!}\int_{(s,t)\in T_{D'}}(1 - s)^{n_1}s^{n_2}ds dt \\
&= 3\frac{n!}{n_1!n_2!}\int_0^{1/2}(1 - s)^{n_1}s^{n_2}\left(\int_{-s}^{s}dt\right)ds \\
&\quad + 3\frac{n!}{n_1!n_2!}\int_{1/2}^{2/3}(1 - s)^{n_1}s^{n_2}\left(\int_{3s-2}^{2-3s}dt\right)ds \\
&= 6\frac{n!}{n_1!n_2!}\left\{4J_{1/2}(n_2 + 2, n_1 + 1)\right. \\
&\quad + 2\left[J_{2/3}(n_2 + 1, n_1 + 1) - J_{1/2}(n_2 + 1, n_1 + 1)\right] \\
&\quad \left. -3J_{2/3}(n_2 + 2, n_1 + 1)\right\}
\end{aligned} \tag{9}$$

where $T_{D'} = \{(s,t): t > 3s - 2,\ t < 2 - 3s,\ t \leq s,\ t \geq -s\}$.

Alternatively, if *BAC* is the true order, then θ_1 and θ_2 are error probabilities subject to constraint C. Again assuming a noninformative prior for θ_1 and θ_2, we have

$$P(\mathbf{n}|BAC) = \frac{n!}{n_1!n_2!}\int_{(\theta_1,\theta_2)\in S_C}\theta_1^{n_1}(1 - \theta_1)^{n_2}\,f(\theta_1,\theta_2)d\theta_1 d\theta_2$$

$$= 6\frac{n!}{n_1!n_2!}\{\tfrac{1}{2}J_{\tfrac{1}{3}}(n_1+1, n_2+2) + J_{\tfrac{1}{2}}(n_1+1, n_2+1) - J_{\tfrac{1}{3}}(n_1+1, n_2+1) - 2J_{\tfrac{1}{2}}(n_1+2, n_2+1) + 2J_{\tfrac{1}{3}}(n_1+2, n_2+1)\} \quad (10)$$

By substituting equations 9 and 10 into equations 7 and 8, one can easily calculate the posterior probability for a given locus order.

Evaluating the Posterior Probability of a Map

Given multiple loci, we often will carry out a sequence of two- and three-color experiments to construct a map of all the loci. We turn now to the problem of evaluating the posterior probability of such a multilocus order given the data. Suppose there are m loci $A_1, A_2,..., A_m$. Suppose L FISH experiments, $E_1, E_2,..., E_L$, are performed, yielding data $\mathbf{n}_1, \mathbf{n}_2,..., \mathbf{n}_L$, where $\mathbf{n}_i = (n_{i1},..., n_{ik_i})$, and $k_i = 2$ or 3, depending on whether two- or three-color data are analyzed. Let S_i be the set of loci ordered in the ith experiment. Let $\mathbf{R}$ be a multilocus order A_{j_1}–A_{j_2}–...–A_{j_m} and let $\mathbf{R} \wedge S$ be the order that conforms to $\mathbf{R}$ but involves only the loci in S. For example, if $\mathbf{R} = A_1$–A_3–A_5–A_4–A_2 and $S = \{A_2, A_3, A_5\}$, $\mathbf{R} \wedge S$ denotes the order A_3–A_5–A_2 or A_2–A_5–A_3. But if $S = \{A_2, A_5\}$, that is, if the data come from a two-color metaphase experiment, $\mathbf{R} \wedge S$ denotes the order A_5–A_2, because the orientation is known. Because orientation does matter for a map, there are total of m possible orders. Suppose the data $\mathbf{n} = (\mathbf{n}_1,..., \mathbf{n}_L)$ suggest a map order:

$$\mathrm{R}_1: \quad A_1 - A_2 - \ldots - A_m$$

We want to calculate the posterior probability $P(\mathrm{R}_1|\mathbf{n})$. By Bayes's theorem,

$$P(\mathrm{R}_1|\mathbf{n}) = \frac{P(\mathbf{n}|\mathrm{R}_1)\,P(\mathrm{R}_1)}{P(\mathbf{n})} = \frac{P(\mathbf{n}|\mathrm{R}_1)\,P(\mathrm{R}_1)}{\sum_{j=1}^{m!} P(\mathbf{n}|\mathrm{R}_j)\,P(\mathrm{R}_j)}$$

Given L experiments $E_1,..., E_L$ and any map order $\mathbf{R}$, in general

$$P(\mathbf{n}|\mathrm{R}_j) = P(\mathbf{n}_1|\mathrm{R}_j)\prod_{i=2}^{L} P(\mathbf{n}_i|\mathrm{R}_j, \mathrm{n}_1, \ldots, n_{i-1})$$

If we make the assumption that, given results $\mathbf{n}_1,...,\mathbf{n}_{i-1}$ obtained through previous experiments $E_1,..., E_{i-1}$, the error probabilities still have a noninformative prior, then $P(\mathbf{n}_i|\mathbf{R}_j, \mathbf{n}_1,..., \mathbf{n}_{i-1}) = P(\mathbf{n}_i|\mathbf{R}_j)$. Obviously, this assumption is correct when loci involved in experiment E_i do not include any of the loci involved in the previous experiments $E_1,..., E_{i-1}$; it is only approximately correct otherwise. However, if the total number of loci is moderate or large, say larger than five, this approximation should be good because the number of possible map orders is so large that modification of the prior owing to previous results will have little impact. Furthermore, this assumption is consistent with the noninformative prior assumption we made in evaluating the posterior probability for a single two- or three-color FISH experiment. In addition, because we are primarily interested in posterior probabilities for those most probable map orders, assuming the noninformative prior throughout the experiments will generally make the posterior probability conservative. This will be true if there is a single order with which all experiments are consistent. In that most important case, were we able to update the prior information at each stage instead of assuming $P(\mathbf{n}_i|\mathbf{R}_1, \mathbf{n}_1,...,\mathbf{n}_{i-1}) = P(\mathbf{n}_i|\mathbf{R}_1)$, the overall most likely order $\mathbf{R}_1$ would be more strongly supported. Thus, the approximation will be conservative. In fact, it is more so when the number of loci is moderate or large, because the information on the map order that is contained in each single experiment diminishes as the number of loci increases. In cases where the data are equivocal, continuing to assume $P(\mathbf{n}_i|\mathbf{R}_1, \mathbf{n}_1,..., \mathbf{n}_{i-1}) = P(\mathbf{n}_i|\mathbf{R}_1)$ will have little impact on our conclusions, because equivocal data would only slightly alter this calculation. In contrast, if the experiments are contradictory, our strategy could result in a posterior probability that is either too large or too small. However, in this situation, it is likely that we would carry out additional experiments in an attempt to rectify the contradictory data. Because $P(\mathbf{R}) = 1/m!$ for all orders $\mathbf{R}$,

Since $P(\mathrm{R}_1|\mathbf{n}) = \frac{1}{m!}$ for all orders $\mathbf{R}$,

$$P(\mathbf{R}_1|\mathbf{n}) \doteq \frac{\prod_{i=1}^{L} P(\mathbf{n}_i|\mathbf{R}_1)\,P(\mathbf{R}_1)}{\sum_{j=1}^{m!}\prod_{i=1}^{L} P(\mathbf{n}_i|\mathbf{R}_j)\,P(\mathbf{R}_j)} = \frac{\prod_{i=1}^{L} P(\mathbf{n}_i|\mathbf{R}_1)}{\sum_{j=1}^{m!}\prod_{i=1}^{L} P(\mathbf{n}_i|\mathbf{R}_j)} \quad (11)$$

Because the ith experiment involves only loci in S_i and, consequently, orders involving other loci are irrelevant to data $\mathbf{n}_i$,

$$P(\mathbf{R}_1|\mathbf{n}) \doteq \frac{\prod_{i-1}^{L} P(\mathbf{n}_i|\mathbf{R}_1 \wedge S_i)}{\sum_{j=1}^{m!}\prod_{i=1}^{L} P(\mathbf{n}_i|\mathbf{R}_j \wedge S_i)} \quad (12)$$

To avoid possible underflow problems, equation 12 can be rewritten as

$$P(\mathbf{R}_1|\mathbf{n}) \doteq \frac{\prod_{i=1}^{L} p_{\mathbf{R}_{1\wedge S_i}}}{\sum_{j=1}^{m!}\prod_{i=1}^{L} p_{\mathbf{R}_{j\wedge S_i}}} \quad (13)$$

where

$$p_{\mathbf{R}_{j\wedge S_i}} = \frac{P(\mathbf{n}_i|\mathbf{R}_j \wedge S_i)}{\sum_{R_{il}\in O_{S_i}} P(\mathbf{n}_i|\mathbf{R}_{il})} \quad (14)$$

and O_{S_i} is the set of k_i orders for the loci in S_i.

Equation 14 can be evaluated as before for two-color, three-color, or two-color–three-locus mapping experiments using equations 2 and 5–10.

Computations

For a modest number of loci m (say $m \leq 12$), it is feasible to enumerate all $m!$ locus orders and hence to identify the

best locus order and to evaluate the denominator of equation 13. However, for larger m, explicit evaluation of all locus orders is impractical. For example, if $m = 14$, >87 billion orders would need to be considered, a daunting task even for a fast workstation. Because for most locus orders, posterior probabilities $P(\mathbf{n}|\mathbf{R})$ are negligibly small, the denominator of equation 13 can be approximated accurately by summing the probabilities of the K most likely locus orders for some K. To identify these best orders when m is large, we use simulated annealing (Kirkpatrick et al. 1983).

Simulated annealing is a technique for combinatorial optimization for problems of very large scale. In simulated annealing, we construct a Markov chain with m states, each state corresponding to a specific locus order. For our implementation of simulated annealing, we have chosen as possible transitions $m(m-1)/2$ two-locus exchanges of the current locus order. For example, if we are in the state corresponding to locus order 4–1–6–2–3–7–5, we may exchange the positions of two loci, 1 and 7, to yield the new order 4–7–6–2–3–1–5. The two-locus exchange seems to perform better in the current context than the block inversion approach used by Press et al. (1989) and Boehnke et al. (1991), in the sense that it finds the best locus orders faster. A useful adjunct is to keep track of the K best orders visited. In implementation, we use a starting temperature of $T = 10{,}000$ and a cooling schedule factor of 0.98. See Press et al. (1989) for a detailed account of the implementation of the procedure.

Although simulated annealing does not guarantee that the best locus orders will be identified, our experience suggests that it generally finds not only the best locus order but also sufficiently many of nearly-best orders that the denominator in equation 13 can be well approximated. In general, we recommend carrying out simulated annealing several times with different starting orders and comparing the results. Our limited experience suggests that $K \cong 100$ to 200 provides an accurate approximation to the denominator of equation 13 and hence of the posterior probabilities for the best locus orders.

Several comments on the evaluation of the posterior probability of a map are in order. First, because

$$\sum_{j=1}^{K}\prod_{i=1}^{L} p_{\mathbf{R}_j\wedge S_i} < \sum_{j=1}^{m!}\prod_{i=1}^{L} p_{\mathbf{R}_j\wedge S_i},$$

approximating the denominator of equation 13 results in slightly inflated posterior probabilities. However, the odds ratios $P(\mathbf{R}_i|\mathbf{n})/P(\mathbf{R}_j|\mathbf{n})$ for locus orders $\mathbf{R}_i$ and $\mathbf{R}_j$ can be easily calculated exactly, because

$$\frac{P(\mathbf{R}_i|\mathbf{n})}{P(\mathbf{R}_j|\mathbf{n})} = \frac{P(\mathbf{n}|\mathbf{R}_i)/\sum_l P(\mathbf{n}|\mathbf{R}_l)}{P(\mathbf{n}|\mathbf{R}_j)/\sum_l P(\mathbf{n}|\mathbf{R}_l)} = \frac{P(\mathbf{n}|\mathbf{R}_i)}{P(\mathbf{n}|\mathbf{R}_j)}$$

where

$$P(\mathbf{n}|\mathbf{R}_j) = \prod_{l=1}^{L} P(n_l|\mathbf{R}_j) = \prod_{l=1}^{L} P(n_l|\mathbf{R}_j\wedge S_l),$$

which can be calculated exactly. In particular, if the two best locus orders $\mathbf{R}_1$ and $\mathbf{R}_2$ are identified, the odds ratio $P(\mathbf{R}_1|\mathbf{n})/P(\mathbf{R}_2|\mathbf{n})$ gives useful information. Second, equation 13 or its approximation can be used to calculate the posterior probability for any locus order whether or not the data are sufficient to suggest a locus order. For example, suppose there are three loci, A, B, and C, and a two-color metaphase experiment yields data $\mathbf{n} = (n_1, n_2)$, which gives $P(AB|\mathbf{n}) = p$, $P(BA|\mathbf{n}) = 1 - p$. By equation 13, $P(ABC|\mathbf{n}) = P(ACB|\mathbf{n}) = P(CAB|\mathbf{n}) = p/3$, and $P(CBA|\mathbf{n}) = P(BCA|\mathbf{n}) = P(BAC|\mathbf{n}) = (1-p)/3$, which agrees with common sense. Third, in situations in which the best locus order is uncertain, information on the K best locus orders, based on the current data, can be used to guide the choice of loci to be included in future experiments.

Strategies for Ordering a Group of Loci

So far, we have focused on the analysis of FISH mapping data. We next turn to issues of efficient experimental design. Given a group of m loci, we can use two- and three-color experiments to build the overall locus order, two or three loci at a time. For obvious reasons, one would like to do so with minimum effort (but see Discussion). For two-color mapping, we propose a bisection strategy to choose the loci to include in each experiment, whereas for three-color mapping, we propose a trisection strategy (Goradia and Lange 1990). In what follows, we assume that each hybridization yields an unequivocal order for a pair or trio of loci.

A Bisection Strategy for Map Construction by Two-color FISH

Consider first a group of m ordered loci labeled 1*, 2*,..., m* in this order, proximal to distal. Suppose a new locus, w*, is to be placed in this group. The bisection strategy requires that the locus w* be compared with locus l* at position $l = \lceil \frac{m+1}{2} \rceil$, roughly the middle locus in the map, where $\lceil x \rceil$ denotes the integer part of x. This comparison will place w* to either proximal or distal to l*. In either case, roughly half of the map can be discarded and the remaining half is bisected again. This process continues until w* is placed in an interval (j*, ($j+1$)*), $j = 1,\ldots, m-1$, left of 1* or right of m*. Figure 2 depicts application of the bisection strategy to order 12 loci.

Using the same argument as for the binary insertion algorithm for sorting (Knuth 1973), it can be shown that the bisection strategy has the best worst-case performance of $\lceil \log_2 m \rceil + 1$ required experiments among all two-locus strategies for ordering w* relative to a map of m loci.

To order a set of m unordered loci, we can use a sequential bisection strategy beginning with any two loci and adding the remaining loci one at a time. The total number $B(m)$ of experiments needed to order m loci using sequential bisection is no more than

$$B(m) = \sum_{k=1}^{m-1} \{\lceil \log_2 k \rceil + 1\}$$

Using summation by parts (Knuth 1973),

$$B(m) = (m-1)\lceil \log_2(m-1) \rceil - \sum_{k=1}^{m-2} k(\lceil \log_2(k+1) \rceil - \lceil \log_2 k \rceil) + (m-1) = m\lceil \log_2(m-1) \rceil - 2^{\lceil \log_2(m-1) \rceil + 1} + m + 1$$

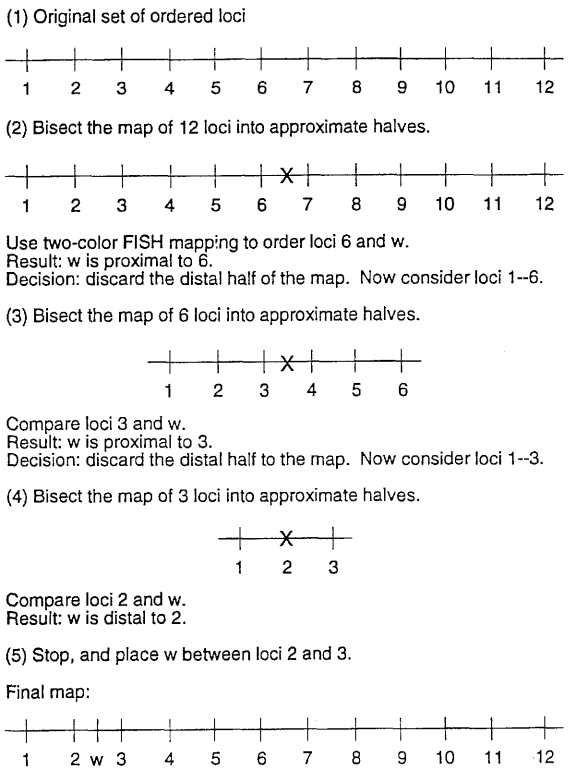


Figure 2 Bisection strategy for adding a new marker, *w*, to a map of 12 loci. The true location of *w* is between 2 and 3.

It also can be shown that the lower bound on the number of experiments among all ordering strategies is $\lceil \log_2 m! \rceil$ (Knuth 1973).

Adapting the argument of Knuth (1973), it easily can be shown that

$$\lim_{m\to\infty} \frac{B(m)}{\lceil \log_2 m! \rceil} = \lim_{m\to\infty} \frac{B(m)}{m \log_2 m} = 1$$

which implies that the sequential bisection strategy for ordering m loci is asymptotically optimal. For any numbers of markers m, this ratio is always ≤ 1.17.

A Trisection Strategy for Map Construction by Three-color FISH

For three-color FISH, we can directly use a trisection strategy proposed by Goradia and Lange (1990) for sperm typing. The trisection strategy is similar to the bisection strategy except that, at each stage, one divides the existing map into three approximately equal blocks instead of two. To place a marker in an existing map $1^*, 2^*, \ldots, m^*$, ordered left to right, we choose two end point loci at positions $\lceil (m+1)/3 \rceil$ and $\lceil (2m+2)/3 \rceil$ and carry out a three-locus FISH mapping experiment using them together with locus w^*. This FISH experiment determines whether w^* falls to the left of $\lceil (m+1)/3 \rceil^*$, between $\lceil (m+1)/3 \rceil^*$ and $\lceil (2m+2)/3 \rceil^*$, or to the right of $\lceil (2m+2)/3 \rceil^*$. Once this decision is made, the chosen interval is then further trisected by combining w^* with loci that are one-third and two-thirds of the way along the interval, and the remaining two-thirds of the map are discarded. This process continues until w^* is placed in a specific interval.

Table 6. Simulation Results for Two-color FISH Experiments for Various Combinations of Error Rate and Sample Size

Error	*n*	Mean $P(R_1\|n)$	$Pr(P(R_1\|n) \geq 0.90)$	$Pr(P(R_1\|n) \geq 0.95)$	$Pr(P(R_1\|n) \geq 0.99)$
0.05	5	0.96	0.77	0.77	0.00
	10	0.99	0.99	0.99	0.91
	20	1.00	1.00	1.00	1.00
0.10	10	0.98	0.93	0.93	0.73
	20	1.00	1.00	1.00	0.96
	30	1.00	1.00	1.00	1.00
0.15	10	0.96	0.82	0.82	0.53
	20	0.99	1.00	0.99	0.83
	30	1.00	1.00	1.00	0.97
0.20	20	0.98	0.97	0.92	0.63
	30	0.99	0.99	0.97	0.87
	40	1.00	1.00	0.99	0.95
0.25	20	0.95	0.90	0.78	0.40
	30	0.98	0.95	0.89	0.67
	40	0.99	0.97	0.95	0.82
	50	1.00	1.00	0.99	0.90

Each row is based on 5000 replications.

Table 7. Simulation Results for Three-color FISH Experiments for Various Combinations of Error Rates and Sample Size

θ_1	θ_2	n	Mean $P(R_1\|n)$	$Pr(P(R_1\|n) \geq 0.90)$	$Pr(P(R_1\|n) \geq 0.95)$	$Pr(P(R_1\|n) \geq 0.99)$
0.05	0.05	10	0.98	0.97	0.93	0.73
	0.05	20	1.00	1.00	1.00	0.99
0.10	0.05	10	0.97	0.91	0.82	0.54
	0.05	20	1.00	1.00	0.99	0.93
	0.05	30	1.00	1.00	1.00	0.99
0.10	0.10	20	0.99	0.99	0.97	0.82
	0.10	25	1.00	1.00	0.99	0.94
	0.10	30	1.00	1.00	1.00	0.98
	0.10	40	1.00	1.00	1.00	1.00
0.15	0.05	20	0.99	0.98	0.94	0.77
	0.05	30	1.00	1.00	0.99	0.95
	0.05	40	1.00	1.00	1.00	0.99
	0.05	50	1.00	1.00	1.00	1.00
0.15	0.10	30	1.00	1.00	0.98	0.90
	0.10	40	1.00	1.00	1.00	0.98
	0.10	50	1.00	1.00	1.00	0.99
	0.10	60	1.00	1.00	1.00	1.00
0.15	0.15	30	0.99	0.98	0.94	0.76
	0.15	40	1.00	1.00	0.99	0.93
	0.15	50	1.00	1.00	1.00	0.97
	0.15	60	1.00	1.00	1.00	0.99
0.20	0.05	20	0.97	0.94	0.84	0.56
	0.05	30	0.99	0.98	0.96	0.82
	0.05	40	1.00	1.00	0.98	0.93
	0.05	50	1.00	1.00	1.00	0.97
	0.05	60	1.00	1.00	1.00	0.99
0.20	0.10	30	0.98	0.96	0.92	0.72
	0.10	40	0.99	0.99	0.97	0.87
	0.10	50	1.00	1.00	0.99	0.95
	0.10	60	1.00	1.00	1.00	0.98

Posterior probalitities are computed under the general error probability model. Each row is based on 5000 replications.

Goradia and Lange (1990) proved that the trisection strategy has the best worst-case performance of $\lceil \log_3 m \rceil + 1$ required experiments among all three-locus strategies for placing a new locus w^* relative to a map of m loci of known order. They also proved that the sequential trisection strategy for ordering m loci requires no more than $T(m)$ experiments, where

$$T(m) = \sum_{k=2}^{m-1} \{\lceil \log_3 k \rceil + 1\}$$
$$= m\lceil \log_3(m-1) \rceil - \frac{3}{2}(3^{\lceil \log_3(m-1) \rceil} - 1) + m - 2$$

The sequential trisection strategy is nearly optimal for any m and is optimal for large m (Goradia and Lange 1990).

Comparison of Bisection and Trisection Strategies

We have shown in the previous sections that bisection and trisection strategies provide the best worst-case performance for placing an additional locus in an existing map and are nearly optimal for building a map. Because metaphase mapping and interphase mapping have different levels of resolution, the choice of which to use may be specified by the distances between the markers. However,

in situations in which either may be used, we can ask, Which is more efficient? To answer this question, let ρ be the cost ratio for conducting a three-color versus a two-color FISH experiment. Comparing worst-case scenarios, the cost in using the three-color scheme relative to the two-color scheme is

$$\frac{\lceil \log_3 m \rceil + 1}{\lceil \log_2 m \rceil + 1} \rho$$

for placing a new marker in an existing map with m loci.

For ordering a set of m loci whose order is completely unknown, it costs no more than $cB(m)$ to build up a map using the two-locus scheme, where c is the unit cost for conducting a two-locus experiment. The corresponding cost for the three-locus scheme is $c\rho T(m)$. To allow for the fact that the three-locus scheme does not provide orientation on the chromosome, suppose that in that case one additional *two-color* experiment is conducted. Then, the cost ratio to order m loci by the three-color scheme relative to the two-color scheme assuming the worst case for each is

$$\frac{\rho T(m) + 1}{B(m)}.$$

Because $x - 1 < \lceil x \rceil \leq x$, it can be readily shown that

$$\begin{aligned} \lim_{m\to\infty} \frac{\rho T(m) + 1}{B(m)} &= \lim_{m\to\infty} \rho \frac{\lceil \log_3 m \rceil + 1}{\lceil \log_2 m \rceil + 1} \\ &= \lim_{m\to\infty} \rho \frac{\lceil \log_3 m \rceil}{\lceil \log_2 m \rceil} \\ &= \rho \log_3 2 \approx 0.63\, \rho \end{aligned}$$

That is, unless the cost ratio ρ is $<1/0.63 \cong 1.58$, there will be no savings using the three-color scheme. Because our experience suggests that three-color interphase FISH is roughly twice as expensive as two-color metaphase FISH, that is, $\rho \cong 2$, use of two-color metaphase mapping appears to be the method of choice, provided resolution is not an issue.

Sample Size Considerations

It is of practical importance to know in advance how many scorings (n) are needed to discriminate between competing locus orders for two- and three-color experiments. In general, n increases with error rates, which are typically unknown in practice. To get a feel of how many scorings are needed, we conducted a simulation study for the two- and three-color experiments for several sample size and error rate combinations. For given error rates and sample size, scorings were simulated via a series of binomial or trinomial experiments, and the posterior probability $P(R_1|n)$ of the correct locus order R_1 was calculated. We repeated this process 5000 times and estimated the average posterior probability and the probabilities of $\Pr[P(R_1|n) \geq 0.90]$, $\Pr]P(R_1|n) \geq 0.95]$, and $\Pr[P(R_1|n) \geq 0.99]$. The results are shown in Tables 6 and 7 for two- and three-color experiments, respectively.

It can be seen from the tables that, for most practical cases where $\theta \leq 0.25$ for two-color experiments and $\max(\theta_1, \theta_2) \leq 0.20$ and $\theta_1 + \theta_2 \leq 0.30$ for three-color experiments, $n = 50$ is usually large enough, whereas for smaller error rates, much smaller sample sizes should be sufficient. To minimize the possibility of choosing the wrong order, we recommend accepting a locus order with a high posterior probability, say ≥ 0.95, or ≥ 0.99.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grant R29GM52205. We are grateful to Dr. Michael Boehnke for his initiation of this project, his technical help, and his critical comments on an earlier version of this paper. We also would like to thank Drs.Kenneth Lange and Soumitra Ghosh for their helpful comments. We also thank two anonymous reviewers for helping to clarify several points in the paper.

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

Alam, K. 1971. On selecting the most probable category. *Technometrics* **13:** 843–850.

Barnes, D.E., K. Kodama, K. Tynan, B.J. Trask, M. Christensen, P.J. De Jong, N.K. Spurr, T. Lindahl, and H.W. Mohrenweiser. 1992. Assignment of the gene encoding DNA ligase I to human chromosome 19q13.2-13.3. *Genomics* **12:** 164–166.

Bechhofer, R.E. and D.M. Goldsman. 1985. On the Ramey-Alam sequential procedure for selecting the multinomial event which has the largest probability. *Comm. Stat. Simul. Comput.* **14:** 263–282.

Berger, J.O. 1985. *Statistical decision theory and Bayesian analysis,* 2nd edition. Springer-Verlag, New York, NY.

Berry, D. 1987. Interim analysis in clinical trials: The role of the likelihood principle. *Am. Stat.* **41:** 117–122.

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

Brandriff, B.F., L.A. Gordon, and B.J. Trask. 1991. DNA sequence mapping by fluorescence in situ hybridization. *Environ. Mol. Mutagen.* **18:** 259–262.

Brandriff, B.F., L.A. Gordon, K.T. Tynan, A.S. Olsen, H.W. Mohrenweiser, A. Fertitta, A.V. Carrano, and B.J. Trask. 1992. Order and genomic distances among members of the carcinoembryonic antigen (CEA) gene family determined by fluorescence in situ hybridization. *Genomics* **12:** 773–779.

Chumakov, I., P. Rigault, S. Guillou, P. Ougen, A. Billaut, G. Guasconi, P. Gervy, I. LeGall, P. Soularue, L. Grinas, et al. 1992. Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359:** 380–387.

Collins, F.S. 1992. Positional cloning: Let's not call it reverse anymore. *Nature Genet.* **1:** 3–6.

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.

Fain, P. 1992. Third International Workshop on human chromosome 17 mapping 1992. *Cytogenet. Cell Genet.* **60:** 177–186.

Flejter, W.L., C.L. Barcroft, S.-W. Guo, E.D. Lynch, M. Boehnke, S. Chandrasekharappa, F.S. Collins, B.L. Weber, and T.W. Glover. 1993. Multicolor FISH mapping with Alu-PCR amplified YAC clone DNA determines the order of markers in the BRCA1 region on chromosome 17q12-q21. *Genomics* **17:** 624–632.

Flory, P.J. 1989. *Statistical mechanics of chain molecules.* Hanser Publishers, New York, NY.

Foote, S., D. Vollrath, A. Hilton, and D.C. Page. 1992. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. *Science* **258:** 60–66.

Goradia, T.M. and K. Lange. 1990. Multilocus ordering strategies based on sperm typing. *Ann. Hum. Genet.* **54:** 49–77.

Hudson, T.J., L.D. Stein, S.S. Gerety, J. Ma, A.B. Castle, J. Silva, D.K. Slonim, R. Baptista, L. Kruglyak, S.H. Xu, et al. 1995. An STS-based map of the human genome. *Science* **270:** 1945–1954.

Kirkpatrick, S., C.D. Gelatt, and M.P. Vecchi. 1983. Optimization by simulated annealing. *Science* **220:** 671–680.

Knuth, D.E. 1973. *The art of computer programming,* Vol. 3. Addison-Wesley, Reading, MA.

Lawrence, J.B., R.H. Singer, and J.H. McNeil. 1990. Interphase and metaphase resolution of different distances within the human dystrophin gene. *Science* **249:** 928–932.

Lebo, R.V., E.D. Lynch, J. Wiegant, K. Moore, M. Trounstine, and M. van der Ploeg. 1991. Multicolor fluorescence in situ hybridization and pulsed field electrophoresis dissect CMT1B gene region. *Hum. Genet.* **88:** 13–20.

Lichter, P., C.J. Tang, K. Call, G. Hermanson, G.A. Evans, D. Housman, and D.C. Ward. 1990. High-resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science* **247:** 64–69.

Olson, M., L. Hood, C. Cantor, and D. Botsein. 1989. A common language for physical mapping of the human genome. *Science* **245:** 1434–1435.

Press, W.H., B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. 1989. *Numerical recipes: The art of scientific computing (FORTRAN version).* Cambridge University Press, Cambridge, UK.

Ramey, J.T. and K. Alam. 1979. A sequential procedure for selecting the most probable multinomial event. *Biometrika* **66:** 171–173.

Trask, B.J. 1991a. DNA sequence localization in metaphase and interphase cells by fluorescence in situ hybridization. *Methods Cell Biol.* **35:** 3–35.

———. 1991b. Fluorescence in situ hybridization: Applications in cytogenetics and gene mapping. *Trends Genet.* **7:** 149–154.

Trask, B.J., H. Massa, S. Kenwrick, and J. Gitschier. 1991a. Mapping of human chromosome Xq28 by two-color fluorescence in situ hybridization of DNA sequences to interphase cell nuclei. *Am. J. Hum. Genet.* **48:** 1–15.

Trask, B.J., G. Van den Engh, M. Christensen, H.F. Massa, J.W. Gray, and M. Van Dilla. 1991b. Characterization of somatic cell hybrids by bivariate flow karyotyping and fluorescence in situ hybridization. *Somat. Cell. Mol. Genet.* **17:** 117–136.

Trask, B.J., H.F. Massa, and M. Burmeister. 1992. Fluorescence in situ hybridization establishes the order cen-DXS28(C7)-DXS67(B24)-DXS68(L1)-tel in human chromosome Xp21.3. *Genomics* **13:** 455–457.

Van den Engh, G., R. Sachs, and B.J. Trask. 1992. Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* **257:** 1410–1412.

Wilke, C.M., S.W. Guo, B.K. Hall, F. Boldog, R.M. Gemmill, S.C. Chandrasekharappa, H.A. Drabkin, and T.W. Glover. 1994. Multicolor FISH mapping of YAC clones in 3p14 and identification of a YAC spanning both FRA3B and the t(3;8) associated with hereditary renal cell carcinoma. *Genomics* **22:** 319–326.

Yokota, H., G. van der Engh, J.E. Hearst, R.K. Sachs, and B.J. Trask. 1995. Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus. *J. Cell Biol.* **130:** 1239–1249.

Received June 7, 1996; accepted in revised form September 6, 1996.