Proc. Natl. Acad. Sci. USA
Vol. 88, pp. 6501–6505, August 1991
Medical Sciences

Error filtration, interference, and the human linkage map

(genetic linkage/gene map/human genome)

D. C. SHIELDS*, A. COLLINS*, K. H. BUETOW†, AND N. E. MORTON*‡

*CRC Research Group in Genetic Epidemiology, Department of Community Medicine, University of Southampton, South Block, Southampton General Hospital, Southampton, SO9 4XY, United Kingdom; and †Division of Population Science, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111

Contributed by N. E. Morton, April 18, 1991

ABSTRACT **Typing error is a major problem in constructing human linkage maps, leading to incorrect orders and inflating map lengths. An error filter is incorporated into multiple pairwise analysis that corrects for inflation of map lengths and improves recovery of the correct order. Multipoint mapping is more sensitive to error, but when its output is adjusted for both error and interference, map lengths are no longer inflated in proportion to the number of loci and are close to those obtained by multiple pairwise analysis.**

A linkage map is essential for isolating disease loci and useful for sequencing the human genome (1). During the past year typing errors have been recognized as a limiting factor (2, 3), as predicted from experimental organisms (4, 5). Phenotypic misclassification simulates crossing-over and thereby inflates map lengths, reduces apparent interference, and supports incorrect orders (6–8). Control by duplicate typing is a counsel of perfection, unfortunately restricted in practice to a few of the most critical results. Selective retyping of apparent recombinants over small distances and double recombinants over moderate distances is helpful but logistically demanding and detects only a proportion of errors. Therefore, statistical filtration to remove the main effects of mistyping would be useful. We have devised such a method and found it to be reliable. The evidence includes theory, analysis of simulated data, and application to chromosomes 1 and 10, from which interference and map lengths have been inferred.

Linkage maps are based on recombination frequencies between pairs of loci (θ), which range between 0 and 0.5. They are converted to map distances (w) in centimorgans (cM) by a mapping function $w = f(\theta)$, which measures the mean number of exchanges in the interval, not all of which are detected as crossing-over. Each exchange is marked by a chiasma that involves two of the four chromatids in a tetrad, and so two chiasmata correspond to 1 exchange per chromatid and, therefore, to 100 cM. Distances in centimorgans are additive, which does not imply that chiasmata occur independently. In fact, one chiasma inhibits others in the neighborhood. This is called (positive chiasma) interference. To account for the effects of interference a number of alternative mapping functions have been proposed (9). Rao's general function $w = f(\theta, p)$ uses a variable p, which ranges between 1 for null interference and 0 for complete interference (10). On present evidence p in man is about 0.35 for both sexes (11). This is remarkable because the distribution of chiasmata differs between the sexes, so that maps must be sex-specific.

There are various approaches to the problem of reducing linkage data to a map. Multiple pairwise analysis [e.g., the program MAP (5)] reduces the data to all possible pairs of loci and combines them appropriately: the significance of each θ is indicated by a logarithm of odds (lod) score, Z (12). This has the advantages that data are easily stored and combined, interference is incorporated, typing errors are damped by inclusion of pairs at moderate and large distances, error filtration has been implemented, and computation for a map of a hundred or more loci is feasible. Although the $n(n-1)/2$ pairs of n loci are not independent, tests of goodness of fit conform remarkably well to χ^2 (13). However, multiple pairwise maps do not extract all the information from multipoint data of unknown linkage phase. (If parental origin of genes A and B in a double heterozygote $AaBb$ is not specified, the phase is said to be unknown.)

To overcome this difficulty multipoint mapping has been developed [e.g., the program CRI-MAP (14)]. However, with this approach, data from different studies are difficult to retrieve and combine, interference is assumed absent, and typing errors have maximal effect because pairs of loci separated by an informative locus do not enter into the likelihood. Multipoint theory has not been developed to account for error or interference, both of which inflate map lengths. A fully efficient analysis that uses gene frequencies such as the program LINKAGE (15) to extract information from untyped parents is feasible for only a small number of loci, especially in pedigrees with several generations of untested parents.

At present both multiple pairwise and multipoint methods have their advocates. Here we shall attempt to make each method work as well as possible under mistyping.

THEORY

Let ε be the mean frequency of undetected mistyping per locus. If a locus is typed independently on two occasions, the probability of discordance is $D = 2\varepsilon(1-\varepsilon)$. This is also the probability that one of two loci be typed correctly and the other be mistyped. If θ is the apparent recombination rate between two loci, each with error frequency ε, the expected value is

$$\theta = D(1-\theta') + \theta'(1-D), \qquad [1]$$

where θ' is the true recombination rate ($0 \le \theta' \le \theta \le 0.5$). When $\theta' = 0.5$, then $\theta = 0.5$.

The algorithm of Morton and Collins (16) for multiple pairwise analysis minimizes the deviance $\Sigma\,(w_{ij} - \hat{w}_{ij})^2 K_{ij}$, where w_{ij} is an observed map distance for a pair of loci (i and j), $\hat{w}_{ij}$ is its expected value, and K_{ij} is the corresponding amount of information. For simplicity of exposition, we suppress the index k = 1, 2, 3 for males, females, and unspecified sex, respectively, over which summation is taken. To incorporate error filtration with computational

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: cM, centimorgan; MLE, maximum likelihood estimate.

‡To whom reprint requests should be addressed.

efficiency and without approximation, we introduce a triangular matrix with elements δ_{ij} for loci i and j defined below. Each iteration is made on error-filtered map intervals μ'_m. Then the expected map distance $\hat{w}'_{ij}$ between i and j is $\Sigma\ \mu'_m$, the summation being taken over all intervals for which $i \leq m < j$. At completion of each cycle error is added to w'_{ij} to give

$$\hat{w}_{ij} = \hat{w}'_{ij} + \delta_{ij}, \qquad [2]$$

and the expected recombination value $\hat{\theta}_{ij}$ is calculated as the inverse Rao mapping function $\theta(\hat{w}_{ij})$. Under the constraint $\delta_{ij} \geq 0$, the estimate of δ_{ij} is $f(\hat{\theta}_{ij}) - f(\hat{\theta}'_{ij})$, where θ_{ij} and θ'_{ij} come from the previous iteration, starting from the observed values of θ_{ij} and

$$\theta'_{ij} = \begin{cases} \dfrac{\theta_{ij} - D}{1 - 2D} & \text{if } \theta_{ij} > D, \\ 0 & \text{otherwise.} \end{cases} \qquad [3]$$

The difference between successive iterations approaches zero as the deviance is minimized.§ The reliability of this procedure has been studied in simulated and real data.

METHODS AND RESULTS

Deterministic Simulation. To test the performance of this error filter, we simulated data (θ and Z) for a map, introducing error by converting the θ' values to θ according to Eq. **1**. Generation of data by this principle is known as deterministic simulation (18). A map of length W with n equally spaced loci has a distance $W/(n - 1)$ between adjacent loci. The true recombination rate between loci with k intervening loci is the inverse mapping function of $(k + 1)W/(n - 1)$. Given the true map length W', the error frequency ε, the number n of loci, the sample size N for each [corresponding to N_{ijk} in Morton and Andrews (5)], and the mapping parameter p, a computer program generates input to MAP for all $n(n - 1)/2$ pairs of loci. For fully informative loci, multipoint analysis operates only on the $n - 1$ adjacent pairs, so that analysis of these by the MAP program measures the sensitivity of multipoint methods to error when loci are fully informative. The data with error may be investigated with or without error filtration, and the multiple pairwise analyses may be carried out with simultaneous estimation of ε. Table 1 gives results for a large map ($W' = 400$) under a supported mapping parameter ($p = 0.35$), a realistic error frequency ($\varepsilon = 0.01$), a large number of loci ($n = 100$), and a moderate sample size for each pair ($N = 40$). The quantity denoted by χ^2 includes only the noncentral component due to deviation from hypothesis, as there is no sampling error in a deterministic simulation.

Neglect of error by MAP inflates the multiple pairwise map by 20 cM, and evidence of a poor fit of the data to the model is given by a χ^2 of 40.14. When the error filter is applied, χ^2 goes down almost to zero, and the true map length is recovered. An attempt to explain the poor fit without error filtration by reducing interference significantly overestimates the mapping parameter p as 0.46, and only slightly improves the fit. This underlines the importance of filtering error before estimating the extent of interference in real data (a mistyped locus generates apparent "recombinants" in two adjacent intervals, contrary to positive interference). The adjacent pairwise (multipoint) map is inflated by 181 cM, about 2 cM per locus. This is seen clearly in Fig. 1, which shows the effect of varying the number of loci. For adjacent pairwise (multipoint) analysis, map length increases linearly with the number of loci, but quickly plateaus for multiple pairwise analysis because of the norming effect of flanking intervals. Multiple pairwise analysis with error filtration recovers the true map length, regardless of the number of loci in the map.

Monte Carlo Simulation. Difficulties in simulating pedigree data to allow a direct comparison between alternative methods may be overcome by sampling from a distribution of gametes. More than 16,000 *Drosophila* have been scored for nine well-spaced loci on the X chromosome (19). The frequencies for the first five loci have been used for Monte Carlo simulation under typing error. This material provides biological realism at the expense of some small uncertainty about map length and mapping parameter. Under the assumption of complete interference, Morgan *et al.* (19) obtained a map length of 36.3 cM for the 5-point map. The true map length under incomplete interference is very slightly longer, and the mapping parameter p is in the neighborhood of 0.35 as for man, mouse, and *Neurospora* (11).

§An error frequency ε is incorporated into the MAP program by including the command $C = \varepsilon$ into the PA control, which also specifies the mapping parameter p; the default value of C is zero. Like p it cannot be iterated directly by the counting algorithm, but the expected lod score Z_E may be calculated for several values of C and/or p in the neighborhood of the maximal Z_E, from which the maximum likelihood estimates and their standard errors are obtained by quadratic interpolation (17).

Table 1. Deterministic simulation

Method	Parameters ε	p	χ^2	W
Adjacent pairwise (multipoint)	(0)	(0.35)	—	580.61
Multiple pairwise	(0)	(0.35)	40.14	420.49
	(0)	0.46	23.25	450.34
	(0.01)	(0.35)	0.08	399.05

Parameters specified by hypothesis are in parentheses, otherwise estimated by maximum likelihood. ε, Error frequency; p, mapping parameter; χ^2, fit; W, estimated map length. Data was simulated with $\varepsilon = 0.01$, $p = 0.35$, $N = 40$, $n = 100$, and $W' = 400$ (see text for explanation).

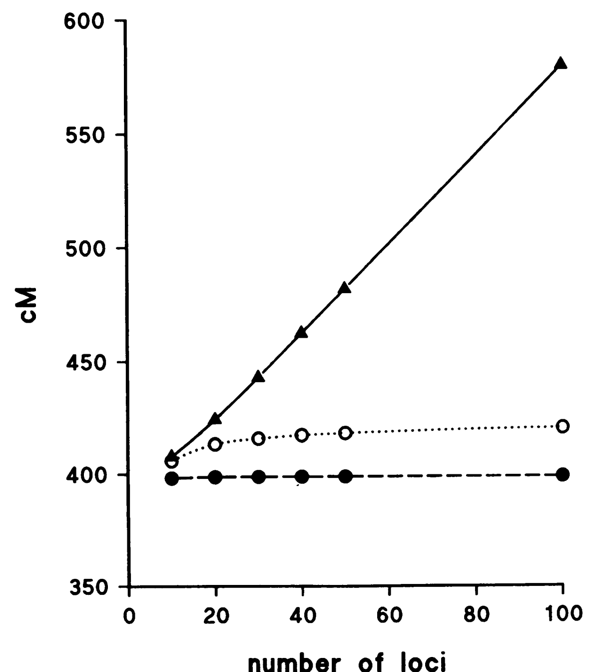

FIG. 1. Effect of number of loci (n) on estimated map length (cM) in a deterministic simulation with error (ε) of 0.01, true map length (W') of 400 cM, and interference parameter (p) of 0.35. ▲, Adjacent pairwise (multipoint) analysis; ○, multiple pairwise analysis, mistyping neglected; ●, multiple pairwise analysis, error filtered.

Table 2. Monte Carlo simulation of *Drosophila* data

Method	Parameters			
	ε	p	χ^2	W
MAP	(0)	(0.35)	1931.38	77.64
	(0)	1.1	690.72	111.80
	0.1	(0.35)	2.49	37.14
	0.1	0.3	1.99	36.64
CRI-MAP	(0)	(0.5)	—	104.4

Data on 25 samples of 40 families (here pooled) were simulated with $\varepsilon = 0.1$, $n = 5$, and $W' = 36.3$. Parameters without parentheses were estimated by maximum likelihood.

With an error frequency of 0.1, multipoint mapping by CRI-MAP gave rise to three wrong orders among 25 simulations. On the assumption of no errors, multiple pairwise mapping gave two wrong orders, one of which was corrected when the error filter was applied. Thus multiple pairwise mapping with error filtration appears to recover the true order more reliably. This experience has been repeated in a much larger series of trials (K.H.B., unpublished data).

With this high error frequency map lengths are markedly overestimated, as 104.4 cM by CRI-MAP and as 77.6 cM by MAP (Table 2). However, with error filtration MAP recovers a realistic map length of 36.6 cM and a mapping parameter of 0.30 and substantially reduces χ^2 (an attempt using MAP to explain the poor fit of the data as interference rather than error only partly improved χ^2, overestimated p, and inflated map length).

Chromosome 10. A difficulty with human linkage data is that there are only approximate *a priori* expectations of true map lengths (based on chiasma frequencies), interference (based on chiasma frequencies and studies in other species), and error (based on duplicate typing of loci). However, using MAP, maximum likelihood estimates of these parameters may be obtained. Table 3 gives results of a MAP analysis of the 28 loci in the Centre d'Etude de Polymorphisme Humain consortium study of chromosome 10 (20), with three megaloci pooled yielding a total of 25 loci. Without error filtration, the maximum likelihood estimate (MLE) of the interference parameter p is larger than expected ($p = 0.6$). With error filtration, there is a significant improvement in fit ($\chi^2_1 = 21.6$). As the joint MLE ($\varepsilon = 0.011$, $p = 0.43$) is not significantly different from the MLE of $\varepsilon = 0.015$ when $p = 0.35$, it seems reasonable to continue to use $p = 0.35$ for chromosome 10, as before (7), and take $\varepsilon = 0.015$. This error filter reduces map length by about 10 cM for the male map and 14 cM for the female map. Thus, the rough error correction we previously applied to chromosome 10 was an overestimate (7). CRI-MAP multipoint analysis gives substantially higher estimates of both male and female map lengths.

To investigate sensitivity to the number of loci, subsets were obtained by systematically removing loci from within the map. Two subsets of 14 loci were obtained by holding the first two and the last locus fixed and removing every second locus. Similarly, three subsets of 18 loci and four subsets of 20 loci were obtained, giving a total of 10 data sets (including the full 25-locus data set). Their MAP error-filtered map lengths ranged between 139.6 and 154.3 cM for males and between 207.5 and 223.7 cM for females. However, there was no significant change in map length with number of loci (Fig. 2). The regression coefficient (or slope, m) for the 10 male data sets was 0.34 (SE = 0.52) and for the female it was −0.47 (SE = 0.62). However, with CRI-MAP map length increases significantly as loci are inserted in the map (Fig. 2). For the 10 male subsets $m = 2.60$ (SE = 0.32) and for the female $m = 2.38$ (SE = 0.35). Thus, multipoint analysis inflates map length by ≈2.5 cM per locus, indicating an error frequency of ≈0.012, which is similar to the MLE of 0.015 obtained by MAP. [m is not exactly 2ε, because it is sensitive to assumptions concerning interference and to the number of matings of unknown phase and because the relationship (Fig. 1) is not exactly linear.]

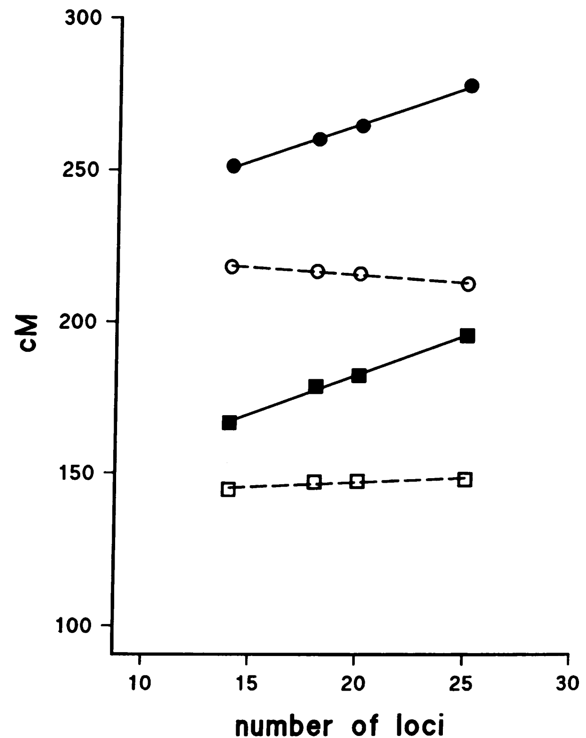

FIG. 2. Effect of number of loci (n) on estimated map length (cM) of chromosome 10 loci. Points represent the mean values for all subsets of a given number of loci (see text). ●, Female map, CRI-MAP multipoint analysis; ○, female map, MAP error filtered ($\varepsilon = 0.015$); ■, male map, CRI-MAP multipoint analysis; □, male map, MAP error filtered ($\varepsilon = 0.015$).

Chromosome 1. The consortium map of chromosome 1 (21) comprises 54 loci (after pooling megaloci). However, it presents some problems: multiple pairwise and multipoint anal-

Table 3. Effects of error and interference on chromosome 10 map lengths

Method	Parameters				
	ε	p	χ^2	W_m	W_f
MAP	(0)	(0.35)	555.25	158.22	228.86
	(0)	0.60	540.04	189.61	275.13
	0.015	(0.35)	533.61	148.01	214.05
	0.011	0.43	532.25	159.14	230.37
CRI-MAP	(0)	(0.5)	—	195.7	277.9

W_m and W_f are male and female distances (cM) for a 25-locus map ranging from D10S31 to D10S6 (20). Parameters without parentheses were estimated by maximum likelihood.

Table 4. Effects of error and interference on chromosome 1 map lengths

Method	Parameters				
	ε	p	χ^2	W_m	W_f
MAP	(0)	(0.35)	2096.24	247.97	423.34
	(0)	0.36	2096.00	249.91	427.04
	0.0070	(0.35)	2061.36	240.80	412.23
	0.0096	0.26	2051.77	222.43	378.08
CRI-MAP	(0)	(0.5)	—	299.5	482.4

W_m and W_f are male and female distances (cM) for a 51-locus map ranging from D1Z2 to D1S68 (21). Parameters without parentheses were estimated by maximum likelihood.

Table 5. Correcting multipoint distances for error and interference

Method	p	Chromosome 10 ♂	Chromosome 10 ♀	Chromosome 1 ♂	Chromosome 1 ♀	*Drosophila* (♀)
CRI-MAP	0.5	195.7	277.9	299.5	485.2	104.5
CRI-MAP, error corrected	0.5	139.3	221.7	242.0	419.9	37.5
CRI-MAP, error corrected	0.35	137.7	217.1	239.5	413.1	37.0
MAP, error filtered	0.35	148.0	214.1	240.8	412.2	37.1

Error correction of CRI-MAP multipoint distances is carried out by converting θ to θ' for each interval estimate (see text). Each θ' is then converted to a distance (cM) using mapping parameter p, which are then summed to give total map lengths. Data sets are identical to those in Tables 2–4.

yses actually disagree in order (21). To eliminate these conflicts, three loci (D1S12, SPTA1, and D1S8) were removed from the map. Results are similar to those for chromosome 10 (Table 4). CRI-MAP multipoint estimates of map length are substantially larger than those of MAP. Error filtration with MAP lowers χ^2 very significantly and reduces male and female map lengths by 7 and 11 cM, respectively. One surprising result is that the low MLE of p (with error filtration) suggests that interference is significantly greater than expected for this data. The MLE of ε when $p = 0.35$ is 0.007. This fits in well with the mean discordance (twice the error rate) of 1.4% obtained by duplicate typing of 14 loci on chromosome 1 (21). The MLE of ε when $p = 0.26$ is slightly higher.

Subsets of the 51 loci were selected in a similar way to the chromosome 10 subsets, giving one data set of 51 loci, four of 39 loci, three of 35 loci, and two of 27 loci. The MAP algorithm failed to converge with one of the female 27 locus data sets after 50 iterations, so it was omitted from the analysis. With MAP the range of map lengths was quite large, between 228.7 and 279.7 for the male data sets and 405.6 and 434.7 for the female data sets. However, this variation was not related to the number of loci in the map ($m = -0.33$, SE $= 0.73$ for the 10 male data sets; $m = -0.07$, SE $= 0.5$ for the 9 female data sets). With CRI-MAP map lengths increased significantly with the number of loci ($m = 0.72$, SE $= 0.41$ for the male; $m = 1.08$, SE $= 0.45$ for the female). Although these results are less conclusive, they suggest an error frequency of $\approx$0.005, which is not inconsistent with MAP's MLE of 0.007.

DISCUSSION

With multipoint methods, an allele typing error of $\approx$1% leads to substantial inflation in map length, rising with the number of loci for both simulated and actual data (Figs. 1 and 2). As linkage maps become more dense, this leads to spurious increases in the map lengths of chromosomes, and genetic locations change with time. Multiple pairwise analysis is less sensitive to error, and an error filter has now been incorporated that protects against inflation in map length.

It is possible to correct the results of multipoint analysis for error to obtain more reliable estimates of map length and gene location. Morton and Andrews (5) proposed as a rough error correction to subtract $200n\varepsilon$ from the total map length. A more exact correction is to take the values of θ for each interval, such as are given in the output of CRI-MAP, and obtain an error corrected value of θ' according to Eq. 3. These recombination fractions may then be converted to a cM scale. To do this, CRI-MAP (version 2.4) uses the Kosambi function (equivalent to $p = 0.5$), which underestimates interference and thereby inflates map length. A more accurate conversion using the Rao function with $p = 0.35$ is given in Table 5. Although these corrections are exact only when all matings are of known phase, they give estimates of map length that are reasonably close to those obtained by MAP for both real and simulated data. Thus, when error and interference are accounted for, multipoint and multiple pairwise estimates of map length agree remarkably well.

Error seems to be greater among the 25 chromosome 10 loci than among the 51 chromosome 1 loci, suggesting that a different error correction should be applied to each chromosome. This is not surprising, as error frequencies vary among loci (21), chromosomes are typed by different laboratories, and the data are subjected to varying levels of scrutiny in the removal of likely typing errors. The error frequency may be estimated either by maximum likelihood using MAP or by examining the increase in map length with number of loci with CRI-MAP. A surprising result of this study is that interference seems to be greater than expected for chromosome 1. It seems reasonable to continue to use the Rao function with $p = 0.35$ for linkage mapping of chromosomes for which there is no reliable estimate of p.

While multipoint estimates of map distance may be corrected for interference and error, the order is obtained without taking error into account. This may lead to incorrect orders when loci are close together. Multiple pairwise analysis reduces the effect of a mistyped locus over a short interval by considering distances over longer intervals, which improves recovery of the correct order. Preliminary results given here indicate that error filtration improves recovery of the correct order still further, but this question needs to be addressed more fully.

We are grateful to Philip Green and Elizabeth R. Robson for comments on this manuscript.

1. Olson, M., Hood, L., Cantor, C. & Botstein, D. (1989) *Science* **245,** 1434–1435.
2. Keats, B. J. B., Sherman, S. L., Morton, N. E., Robson, E. B., Buetow, K. H., Cartwright, P. E., Chakravarti, A., Francke, U., Green, P. P. & Ott, J. (1991) *Genomics* **9,** 557–560.
3. Keats, B. J. B., Sherman, S. L., Morton, N. E., Robson, E. B., Buetow, K. H., Cartwright, P. E., Chakravarti, A., Francke, U., Green, P. P. & Ott, J. (1991) *Ann. Hum. Genet.* **55,** 1–6.
4. Lane, P. W. (1963) *J. Hered.* **54,** 263–266.
5. Morton, N. E. & Andrews, V. (1989) *Ann. Hum. Genet.* **53,** 263–269.
6. Jiang, O. X. & Buetow, K. H. (1990) *Am. J. Hum. Genet.* **47** (Suppl.), A185 (abstr.).
7. Morton, N. E. & Collins, A. (1990) *Ann. Hum. Genet.* **54,** 235–251.
8. Terwilliger, J. D., Weeks, D. E. & Ott, J. (1990) *Am. J. Hum. Genet.* **47** (Suppl.), A201 (abstr.).
9. Liberman, U. & Karlin, S. (1984) *Theor. Pop. Biol.* **25,** 331–346.
10. Rao, D. C., Morton, N. E., Lindsten, J., Hulten, M. & Yee, S. (1977) *Hum. Hered.* **27,** 99–104.
11. Morton, N. E., MacLean, C. J. & Lew, R. (1985) *Genet. Res.* **45,** 279–286.
12. Morton, N. E. (1955) *Am. J. Hum. Genet.* **7,** 277–318.
13. Rao, D. C., Keats, B. J., Lalouel, J.-M., Morton, N. E. & Yee, S. (1979) *Am. J. Hum. Genet.* **31,** 680–696.
14. Lander, E. S. & Green, P. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2363–2367.
15. Lathrop, G. M., Lalouel, J.-M., Julier, L. & Ott, J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3443–3446.
16. Morton, N. E. & Collins, A. (1990) *Ann. Hum. Genet.* **54,** 103–106.

17. Ott, J. (1985) *Analysis of Human Linkage Data* (Johns Hopkins Univ. Press, Baltimore), pp. 197–199.
18. Morton, N. E. (1984) in *Human Population Genetics*, ed. Chakravarti, A. (Von Nostrand Reinhold, New York), pp. 83–107.
19. Morgan, T. H., Bridges, C. B., Schultz, J. (1935) *Carnegie Inst. Washington Yearb.* **34**, 284–291.
20. White, R. L., Lalouel, J. M., Nakamura, Y., Donis-Keller, H., Green, P., Bowden, D. W., Mathew, C. G. P., Easton, D. F., Robson, E. B., Morton, N. E., Gusella, J. F., Haines, J. L., Retief, A. E., Kidd, K. K., Murray, J. C., Lathrop, G. M. & Cann, H. M. (1990) *Genomics* **6**, 393–412.
21. Dracopoli, N. C., O'Connell, P., Elsner, T. I., Lalouel, J. M., White, R. L., Buetow, K. H., Nishimura, D. Y., Murray, J. C., Helms, C., Mishra, S. K., Donis-Keller, H., Hall, J. M., Lee, M. K., King, M. C., Attwood, J., Morton, N. E., Robson, E. B., Mahtani, M., Willard, H. F., Royle, N. J., Patel, I., Jeffreys, A. J., Verga, V., Jenkins, T., Weber, J. L., Mitchell, A. L. & Bale, A. E. (1991) *Genomics* **9**, 686–700.