GENOMICS 14, 604–610 (1992)

Systematic Detection of Errors in Genetic Linkage Data

STEPHEN E. LINCOLN AND ERIC S. LANDER

Whitehead Institute for Biomedical Research and Department of Biology and Center for Genome Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 01242

Received March 9, 1992; revised July 17, 1992

Construction of dense genetic linkage maps is hampered, in practice, by the occurrence of laboratory typing errors. Even relatively low error rates cause substantial map expansion and interfere with the determination of correct genetic order. Here, we describe a systematic method for overcoming these difficulties, based on incorporating the possibility of error into the usual likelihood model for linkage analysis. Using this approach, it is possible to construct genetic maps allowing for error and to identify the typings most likely to be in error. The method has been implemented for F2 intercrosses between two inbred strains, a situation relevant to the construction of genetic maps in experimental organisms. Tests involving both simulated and real data are presented, showing that the method detects the vast majority of errors. © 1992 Academic Press, Inc.

INTRODUCTION

Genetic linkage maps consisting of DNA polymorphisms are powerful tools for the study of inheritance, facilitating chromosomal localization and eventual cloning of genes causing traits and diseases (Botstein *et al.*, 1980). Because of their great utility, such genetic maps are being constructed in many organisms. In some cases, including the human and the mouse, preliminary maps containing hundreds of markers have already been developed (Donis-Keller *et al.*, 1987; Copeland and Jenkins, 1991; Dietrich *et al.*, 1992) and attention is now turning to the construction of dense genetic maps with markers spaced every centimorgan—which are useful for many purposes, including fine-scale genetic mapping and top-down anchoring of physical maps.

In principle, construction of increasingly dense genetic maps is a straightforward extension of existing work. In practice, however, such efforts confront a major obstacle: laboratory typing error. Given the large number of typings required to construct a map, some nonzero error rate seems inevitable, and recent estimates suggest that the actual error rates may be in the range 0.5–3.0% (Dracopoli *et al.*, 1991; Patterson, 1991; Dietrich *et al.*, 1992). Whereas such an error rate has only slight effects on a sparse genetic map, it can have disastrous effects as the map grows denser. The reason is not hard to understand: Errors tend to introduce spurious crossovers—increasing the apparent genetic length of intervals and decreasing the support for the correct genetic order (Ott, 1977; Buetow, 1991; Morton, 1991; Chakravarti and Lasher, 1992). As the map gets denser and recombination frequency between markers approaches the error rate, a significant proportion of all observed crossovers will be spurious.

A simple first-order approximation shows the magnitude of the problem: With an error rate of $\epsilon = \delta/100$, a small interval of true size d cM will be inflated to an apparent size of $d_{\text{inf}} = d + 2c\delta$ cM, where c is a constant depending on the nature of the cross. The estimate is strictly valid only for small intervals and somewhat overestimates the expansion for larger intervals. Nonetheless, it makes clear that there will be substantial map expansion: a genetic map of true length L cM will be inflated to an apparent length of about $L_{\text{inf}} = L + 2c\delta N$ cM, where N is the number of intervals studied. In short, as more markers are added, the apparent genetic length of the map grows, and the goal of a dense map continually recedes.

The estimate is based on the following approximations, which are correct up to second-order terms in d and ϵ: (1) the apparent size in centimorgans of a small interval is equal to the apparent recombination frequency in percent; (2) the apparent recombination frequency of a small interval is approximately the proportion of true crossovers plus spurious crossovers; and (3) the proportion of the spurious crossovers is approximately $2\epsilon c$, where 2ϵ is approximately the probability that exactly one of the flanking markers will be mistyped in a given individual, and c is the expected number of spurious crossovers produced by such a mistyping divided by the expected number of informative meioses per individual. The approximation neglects the second-order effects of (1) the mapping function, (2) the small proportion of cases in which errors eliminate a true crossover, and (3) the small proportion of cases in which both flanking markers are mistyped, but the approximation is nonetheless quite good. The constant c depends on the

0888-7543/92 $5.00

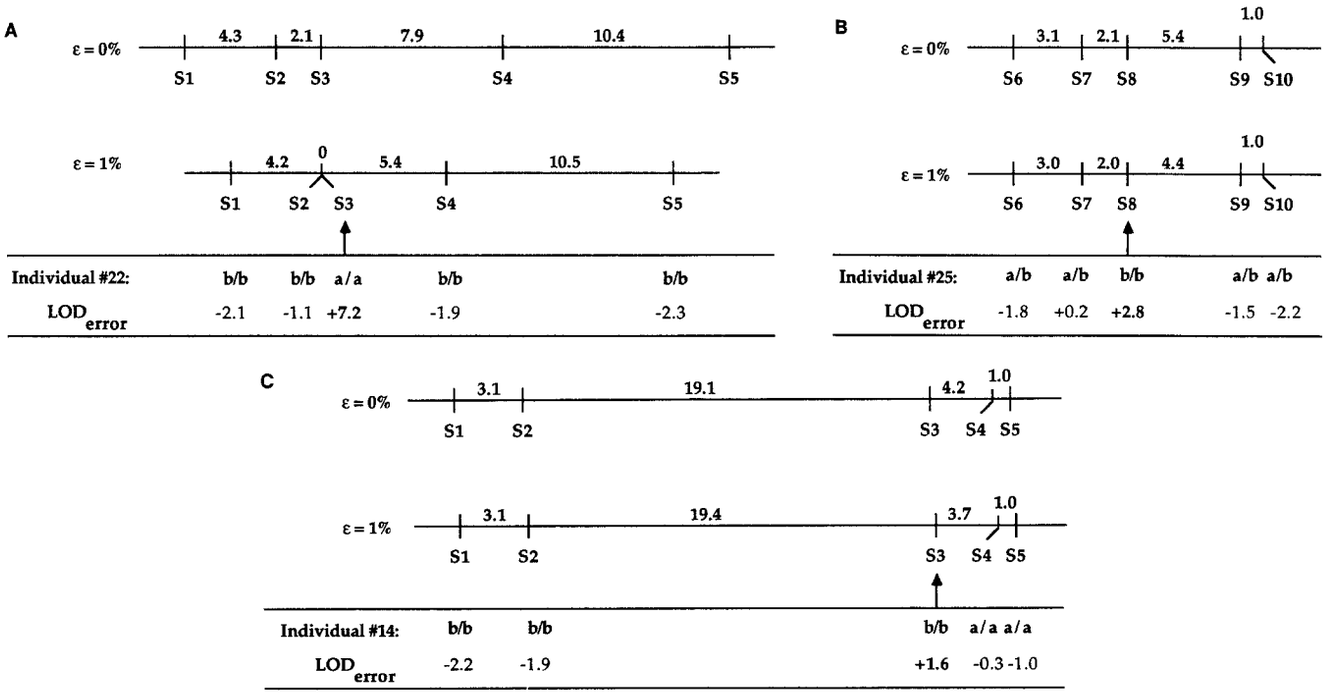

FIG. 1. Three examples illustrating the effects of errors on the construction of genetic maps in actual data from crosses. (**A, B, C**) The map given by the traditional approach is on top and the map given by the error-tolerant approach in the middle. Typings for specific individuals likely to contain an error are at the bottom, with the LOD_{error} for each typing. In each case, an apparently erroneous typing at one locus (shown with an arrow) has occurred. The traditional approach results in map expansion, while the error-tolerant approach is robust to error. Note that, in **C**, the error-tolerant approach shortens the overall map but increases the interval between S2 and S3 because a crossover has almost certainly occurred on both chromosomes in individual 14, but there is a substantial probability that the typing at S3 is erroneous and that at least one crossover occurred in the interval S2–S3.

structure of the cross, the informativeness of the markers, and the nature of the errors. Suppose, for example, that any genotype that is incorrect but consistent with Mendelian inheritance is equally likely to result from mistyping. By simple enumeration of cases, one can show that $c = 1$ for a fully informative backcross (aa/bb × bb/bb); $c = \frac{5}{8}$ for a fully informative intercross (aa/bb × aa/bb); and $c = \frac{2}{3}$ for a fully informative cross (aa/bb × cc/dd).

The problem of map expansion due to the addition of markers is nicely illustrated in the Report of the Second International Workshop on Chromosome 21 (Patterson, 1991). The report compared two genetic maps made in different collections of families but covering precisely the same region. One map contained 12 subintervals and showed a total length of 87 cM, while the other map contained 18 subintervals and showed a total length of 122 cM. The report states, "This increase in map length as more segments are mapped between two framework markers . . . is indicative of undetected genotypic errors. A 50% increase in mapped segments led to a 40% increase in map length." Using our first-order approximation above, such an increase would be consistent with an error rate of about $(122\text{-}87)/((18\text{-}12) \times 2c) = 3c^{-1}\%$. Assuming that the degree of informativeness most closely resembles a backcross ($c = 1$), this would be about 3%. Thus, even small error rates of a few percent can seriously distort the map. Clearly, the construction of dense genetic maps requires methods for overcoming the effect of typing errors.

Here, we present a systematic mathematical method that (i) can construct accurate genetic maps notwithstanding errors and (ii) can identify the most likely errors. The method is based on the traditional maximum likelihood approach for genetic mapping, modified to include explicitly the possibility of laboratory error. We have implemented this method in a computer program for the case of an F2 intercross in an experimental organism. We present results both for simulated data and for actual data from the construction of a dense genetic map in the mouse.

METHODS

Our approach is based on the notion that laboratory typing results should be considered *phenotypes* rather than *genotypes*. To be sure, the true genotypes correspond *very closely* to the phenotype observed in the DNA typing assays, but they do not correspond *perfectly* due to laboratory errors. In other words, the genotypes are incompletely penetrant to a small degree. Fortunately, incomplete penetrance is readily handled within the traditional maximum likelihood approach for genetic mapping: one simply specifies a penetrance function defining the probability of observing each phenotype given the true genotype (see, e.g., Ott, 1985). Accordingly, our method involves (i) specifying a penetrance function based on an assumed error rate, (ii) constructing the most likely genetic map using this penetrance function, and (iii) identifying the likely errors by finding those data for which the probabilistically inferred genotype does not correspond to the observed typing. Below, ϵ^* denotes the true error rate, and ϵ denotes the error rate assumed in the analysis.

Specifically, let X_{ij} denote the observed typing result and G_{ij} denote the true genotype (not directly observed) for individual i at locus j. Given an appropriate penetrance function to describe the effect of

laboratory error, we calculate (1) the maximum likelihood genetic map consisting of a vector of recombination fractions $\hat{\theta} = (\theta_k)$, given the complete typing data $X = \{X_{ij}\}$ and an assumed error rate ϵ; (2) the a posteriori probability distribution for each true genotype G_{ij} given X, ϵ, and $\hat{\theta}$; and then (3) calculate the LOD score for error, defined as

$$\mathrm{LOD}_{\mathrm{error}}(i,j) := \log_{10}\left[\frac{\mathrm{Prob}(X \mid X_{ij} \neq G_{ij}, \hat{\theta} \text{ and } \epsilon)}{\mathrm{Prob}(X \mid X_{ij} = G_{ij}, \hat{\theta} \text{ and } \epsilon)}\right] \quad [1a]$$

$$= \log_{10}\left[\frac{\mathrm{Prob}(G_{ij} \neq X_{ij} \mid X, \hat{\theta} \text{ and } \epsilon)}{\mathrm{Prob}(G_{ij} = X_{ij} \mid X, \hat{\theta} \text{ and } \epsilon)} \Big/ \frac{\epsilon}{1-\epsilon}\right]. \quad [1b]$$

The LOD score for error, $\mathrm{LOD}_{\mathrm{error}}(i,j)$, is the logarithm of the odds ratio of the probability that the complete data set would arise if the typing X_{ij} is incorrect divided by the probability that it would arise if X_{ij} is correct. This can be computed as the ratio of the a posteriori odds ratio in favor of error divided by the ratio a priori odds ratio in favor of error, as shown in (1b). Note that the calculation of $\mathrm{LOD}_{\mathrm{error}}(i,j)$ for the typing for individual i at locus j does *not* assume that the rest of the data are necessarily correct, but rather allows simultaneous search for errors throughout the data. Those typing results X_{ij} with the highest $\mathrm{LOD}_{\mathrm{error}}$ are the most likely "potential errors," given the data. (N.B. The statistic $\mathrm{LOD}_{\mathrm{error}}$, which applies to individual typing, should be carefully distinguished from the statistic LOD, which applies to an entire pedigree. The symbol $\mathrm{LOD}_{\mathrm{error}}$ should always be written out completely.)

In this paper, we consider the case of an F2 intercross between two inbred strains, a situation relevant to the construction of the genetic map in most experimental organisms, including the mouse, rat, and maize. Recall that an F2 intercross involves breeding two inbred strains A and B to produce F1 individuals heterozygous at all loci, which are then crossed to produce F2 progeny. There are three possible genotypes: AA, AB, and BB. For simplicity, we will assume that mistyping occurs with probability ϵ and is equally likely to result in either of the other two possible results. Thus, laboratory typing for each locus obeys the simple penetrance function,

$$\begin{aligned}\mathrm{Prob}(\text{Observed typing} = X \mid \text{True genotype} = G) \\ &= 1 - \epsilon \quad \text{if } X = G \\ &= \epsilon/2 \quad \text{if } X \neq G, \end{aligned} \quad [2]$$

where ϵ is the assumed error rate. The assumed error rate ϵ will be taken to be a fixed parameter, equal for all loci. In principle, ϵ could be specified separately for each locus or even regarded as an unknown parameter to be estimated. In fact, we note below that the error detection is quite insensitive to the choice of assumed error rate, within the range of interest. For this purpose, it is enough to use an estimate based on repeating a fraction of the typings.

As noted above, $c = \frac{5}{8}$ for an F2 intercross with errors occurring as in [2]. Based on our first-order approximation above, we would expect that a genetic map of length L cM would be inflated to $L_{\mathrm{inf}} = L + 125\epsilon N$ by the traditional approach, where N is the number of intervals in the map.

To calculate $\mathrm{LOD}_{\mathrm{error}}$, one simply uses the penetrance function given in (2), performs the usual linkage analysis, and calculates the a posteriori probability distribution over genotypes required in (1b). The a posteriori probability distribution is easily obtained from any of the usual algorithms for linkage analysis. In our case, we adapted the MAPMAKER computer package (Lander *et al.*, 1987) to perform the required calculations for an F2 intercrosses. Since the Hidden Markov chain algorithm used in MAPMAKER explicitly computes the required a posteriori probability distribution (see Lander and Green, 1987), this change only involved substituting the penetrance function. Calculation of $\mathrm{LOD}_{\mathrm{error}}$ for all typings thus did not substantially increase computation time. The modified program is available from the authors.

We examined both simulated and real data. The simulated data were produced as follows: genetic markers were randomly positioned (according to a Poisson process) along a single chromosome of 100 cM with mean spacing of d cM; genotypes were generated for n F2 progeny following standard Mendelian segregation (assuming no crossover interference); and typing results were generated by assuming that errors were independently and identically distributed with probability ϵ. The real data came from a genetic map of the mouse recently constructed in our laboratory (Dietrich *et al.*, 1992). Briefly, the map involves 317 simple sequence length polymorphisms (SSLPs) studied in an F2 intercross between C57BL/6J-ob/ob and CAST/Ei with 46 progeny. The map cover all 20 mouse chromosomes with an average spacing of 4.3 cM and is based on a total of 14,285 typings. By repeating about 10% of the typings, we estimated that the initial error rate was about 0.7%.

RESULTS

Random errors in a dense genetic map tend to produce events that are statistically unlikely under the usual genetic model—such as double crossovers in adjacent intervals or single crossovers in a small interval in both paternal and maternal meiosis. Our error-tolerant method involves constructing the genetic map under a model that allows for both genetic recombination and typing error. As a result, the model can recognize when an event is more likely to be the result of error than recombination. The approach is illustrated in Fig. 1, which shows examples of genetic maps constructed with and without allowance for error, together with the values of $\mathrm{LOD}_{\mathrm{error}}$ for particular individuals with probable errors. Because the method can attribute unlikely events to probable error, it both avoids map expansion and identifies the typings that should be rechecked.

Error detection works well for the internal loci in a map—i.e., those loci with flanking markers on each side—because errors can usually be distinguished from recombination events by examining the flanking markers. By contrast, one has much less statistical power to detect errors at the two terminal loci in a linkage group. Essentially, it is only feasible to detect errors at terminal loci when the terminal interval is quite small and when the error has introduced apparent crossovers in both paternal and maternal meiosis. Because the vast majority

TABLE 1

Average Map Length for Simulated Data with Average Error Rate of 1%, with Maps Constructed by Traditional Method or by Error-Tolerant Method

Average spacing (cM)	Average map length (cM)		
	Actual[a]	Traditional[b]	Error-tolerant[c]
1	98	216	99
2	96	157	98
4	92	120	93
8	87	94	82

[a] Actual length between two most distal markers in map.

[b] Length of map constructed by traditional method ignoring possibility of error.

[c] Length of map constructed by error-tolerant method assuming 1% error rate.

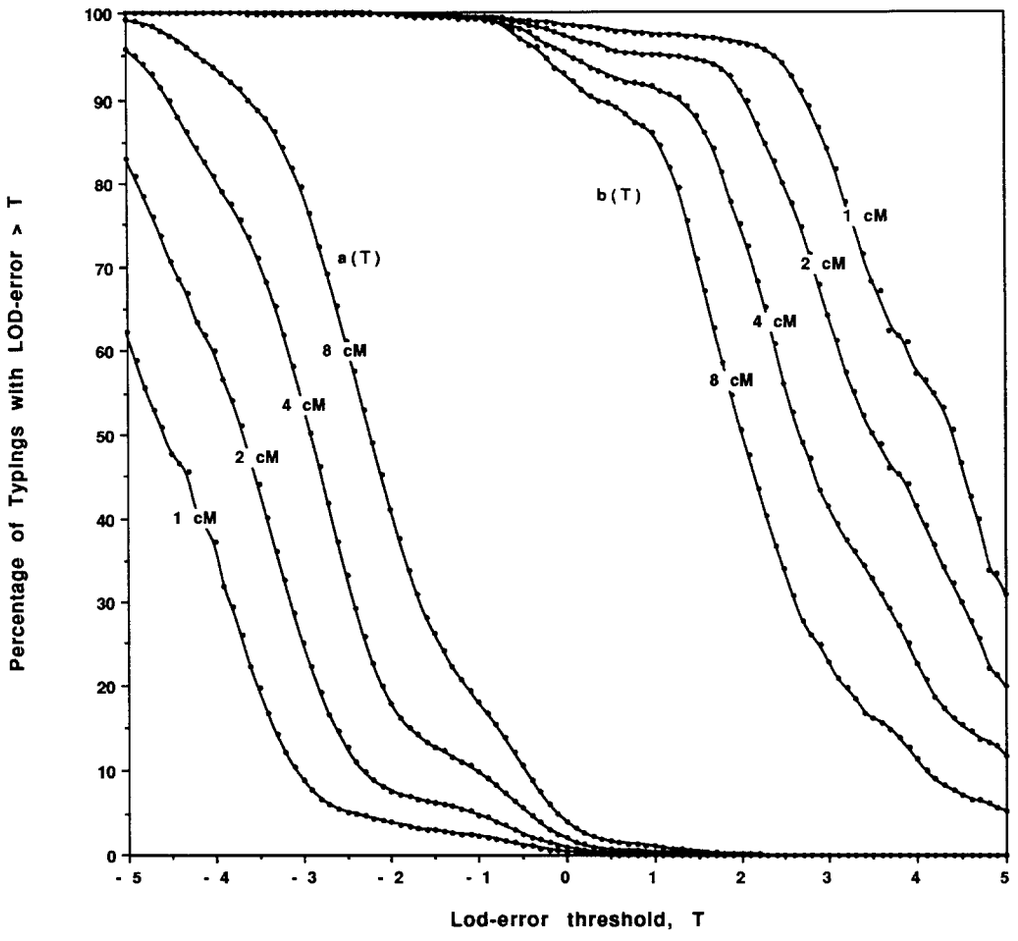

FIG. 2. LOD_{error} for simulations with average spacing of 1, 2, 4, and 8 cM; actual error rate $\epsilon^* = 1\%$ and assumed error rate $\epsilon = 1\%$. Curves show the inverse cumulative probability distributions $a(T)$ = proportion of *correct* typings with $LOD_{error} > T$ and $b(T)$ = proportion of *incorrect* typings with $LOD_{error} > T$. Note that $a(T)$ and $b(T)$ move in opposite directions as the density of the map increases, reflecting the increasing ability to distinguish between correct and incorrect typings using the statistic LOD_{error}.

of loci in a dense map are internal, we focus below exclusively on detecting errors at internal loci.

Simulated Data

We simulated F2 intercrosses involving 100 progeny, with markers randomly distributed with an average spacing of d = 1, 2, 4, and 8 cM along a single chromosome of 100 cM and an actual error rate of 1% in typing (i.e., $\epsilon^* = 0.01$). A total of 100 such simulations were performed for each average spacing.

For each simulation, genetic maps were constructed in two ways: the traditional approach in which the data are assumed to be completely correct ($\epsilon = 0$) and our error-tolerant method assuming that the data have a 1% error rate ($\epsilon = 0.01$). The traditional approach caused substantial map expansion in agreement with our estimates above, while the error-tolerant method gave maps with approximately the correct length (Table 1). Considering the simulation with 4 cM average spacing, we see that the average distance of 92 cM between the most distal markers was estimated as 120 cM based on the traditional approach ignoring the effect of errors, and 93 cM based on our error-tolerant method. The inflation seen in the traditional analysis agrees well with our first-order approximation $L_{inf} = L + 125\epsilon N$ for an intercross, which predicts an apparent length of 121 cM based on about 23 intervals in the map.

We then examined the power of the statistic LOD_{error} to detect errors at internal loci in the map. Specifically, we calculated the inverse cumulative distributions:

$a(T)$ = proportion of *correct* typings with $LOD_{error} > T$;
$b(T)$ = proportion of *incorrect* typings with $LOD_{error} > T$; and
$c(T)$ = proportion of *all* typings with $LOD_{error} > T$.

Note that $c(T) \approx (1 - \epsilon^*)a(T) + \epsilon^* b(T)$.

The distributions $a(T)$ and $b(T)$ are illustrated in Fig. 2 for maps with average spacing of 1, 2, 4, and 8 cM. As the figure makes clear, correct typings tend to have a low value of LOD_{error}, while incorrect typings tend to have a high value—with the tendency becoming more pronounced as the map density increases. Consider, for example, the case of maps with an average density of 4 cM. The set of typings with $LOD_{error} \geq 0$ contains 3% of the total data but 95% of the errors, while the set of typings with $LOD_{error} \geq 1$ contains about 1.3% of the total data but 91% of the errors. Clearly, it is necessary to recheck

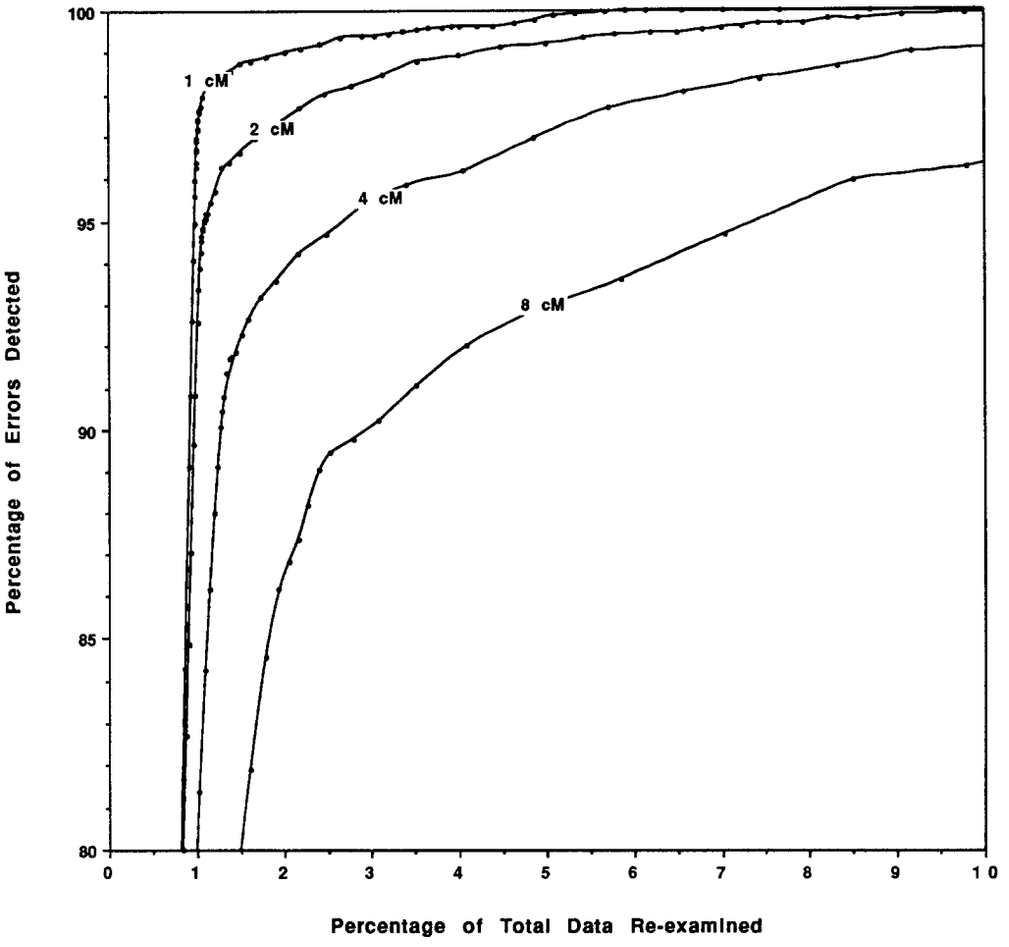

FIG. 3. Trade-off between the proportion of data requiring rechecking and the proportion of errors caught, for simulated maps with average spacing of 1, 2, 4, and 8 cM and actual error rate of $\epsilon^* = 1\%$. Specifically, the curves plot $c(T)$ against $b(T)$ as the threshold T increases. For a different error rate ϵ^{**}, the corresponding graph can be approximately obtained by translating the curves to the right by ϵ^{**}-ϵ^*. This follows from the fact that the distributions $a(T)$ and $b(T)$ are relatively insensitive to the value of ϵ^* and that the distribution $c(T)$ corresponding to an error rate ϵ^* is given by $c(T) = (1 - \epsilon^*)a(T) + \epsilon^* b(T)$.

only a small proportion of the typings to catch the vast majority of the errors. In practice, the proportion of typings to be rechecked should depend on the cost of rechecking data versus the consequences of leaving errors uncorrected.

This trade-off between the proportion of typings requiring rechecking and the proportion of errors caught —i.e., between $c(T)$ and $b(T)$—is shown for maps with average spacing of 1, 2, 4, and 8 cM in Fig. 3. As the map becomes denser, error detection becomes increasingly efficient. By rechecking the highest 1% of the $\text{LOD}_{\text{error}}$ scores, one can detect about 79% of the errors in a 4-cM map, 92% of the errors in a 2-cM map, and 96% of the errors in a 1-cM map. By rechecking the highest 2% of the $\text{LOD}_{\text{error}}$ scores, the proportions of errors caught are 94, 97, and 99%.

In the analyses above, we used a 1% error rate in the simulations and assumed a 1% error rate when performing the analysis. We also explored the consequences of varying the true error rate used in the simulations and the assumed error rate used in the analysis. In fact, the distributions $a(T)$ and $b(T)$ do not change substantially as the true and assumed error rate vary over the range 0.2–4% (data not shown). This stands to reason because error detection depends primarily on observed typing differences between a locus and its neighbors, with ϵ and ϵ^* having only a second-order effect in the calculation.

As a consequence, error detection does not depend sensitively on the assumed error rate ϵ, at least within the range of interest. The most likely errors will have the highest values of $\text{LOD}_{\text{error}}$, regardless of the precise choice of ϵ. It thus suffices to use a rough estimate based on repeating a proportion of the typings.

However, the values of ϵ and ϵ^* do affect the analysis in two ways. The trade-off curves in Fig. 3 need to be modified for different values of the true error rate, because the distribution $c(T)$ changes with ϵ^* according to the formula $c(T) \approx (1 - \epsilon^*)a(T) + \epsilon^* b(T)$ (see legend to Fig. 3). Also, if the assumed rate $\epsilon \neq \epsilon^*$, the map distances produced by the error-tolerant linkage analysis will be statistically biased because too much or too little weight will be placed on observed crossovers.

Weighing these considerations, we recommend using the method as follows: (1) Obtain a rough estimate ϵ of the error rate by repeating a proportion of the typings; (2) using this value of ϵ, identify the typings with a high

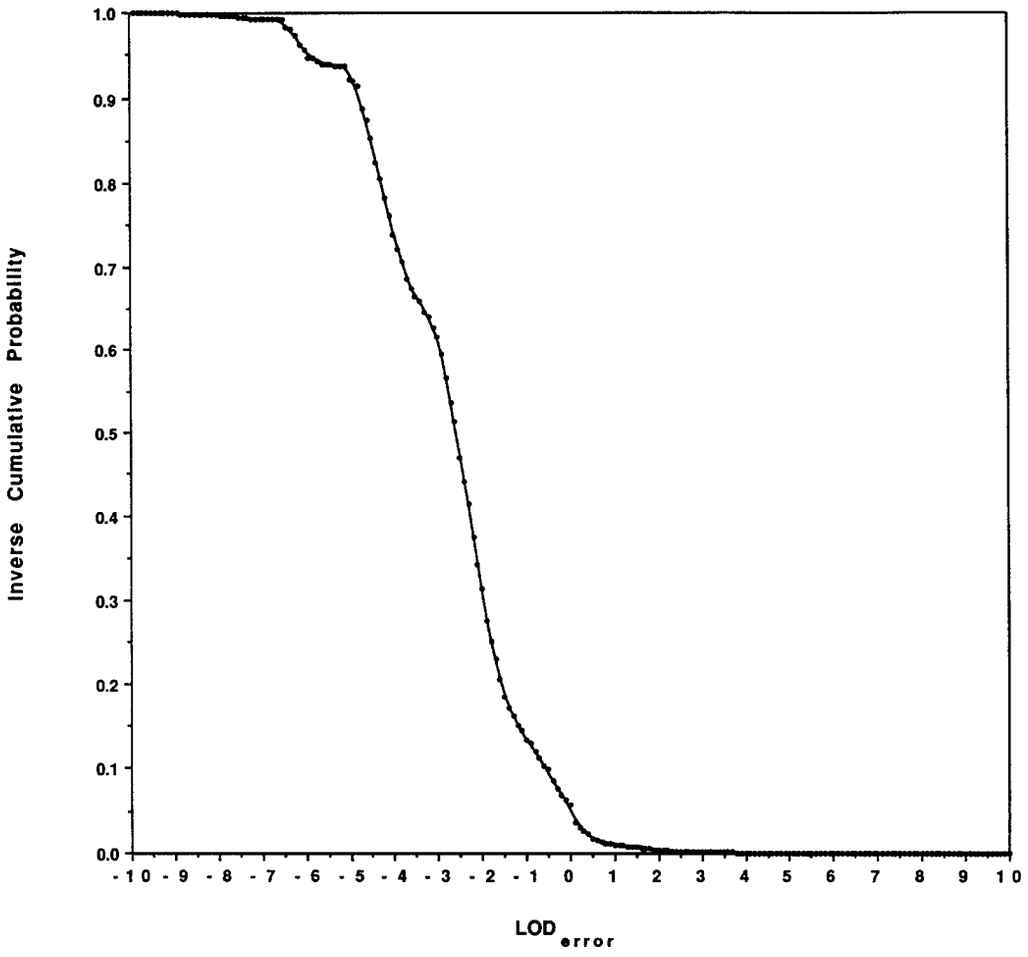

FIG. 4. Inverse cumulative probability of LOD_{error} for dataset underlying 317-marker mouse genetic map (Dietrich *et al.*, 1992).

value of LOD_{error}; (3) recheck these typings, thereby cleaning the data set, reestimating the initial error rate ϵ, and estimating the residual error rate ϵ_{res}; and (4) construct the final map assuming the residual error rate. If the error rate is large, it may be prudent to iterate steps (1)–(3).

Real Data

After studying these simulations, we analyzed actual data from our genetic map of the mouse (Dietrich *et al.*, 1992). Our error rate was initially estimated to be about $\epsilon = 0.7\%$, based on repeating a proportion of the typings. Using this estimate, we constructed the genetic map allowing for error and calculated LOD_{error} for each of the 14,285 individual typings (Fig. 4). About 1.1% of the typings at internal loci had $LOD_{error} > 1$, and these were identified as potential errors. Of these 157 typings, 72 proved to be actual errors upon rechecking. These results accord well with expectation. Specifically, (1) the proportion of potential errors that proved to be actual errors (46%) agreed well with expectation (50%) based on simulations for a map of this spacing and error rate, and (2) the proportion of detected errors in the data set (0.5%) agreed well with expectation (0.6%) based on the estimated error rate of 0.7% and the simulation studies indicating that about 88% of actual errors would be detected (data not shown). Based on the results above, the number of errors at internal loci escaping detection might be about 10 of the 14,285 typings. In addition, one might expect another 10 errors at terminal loci in the map. In fact, we rechecked all terminal loci but detected no errors.

Interestingly, the detected errors were not independently distributed across loci. There were 280 loci with no errors, 23 loci with one error, 3 loci with two errors, 3 loci with three errors, 4 loci with four errors, 1 locus with five errors, and 2 loci with six errors. The handful of loci with many errors were the SSLPs with patterns that were especially difficult to interpret.

As expected, the genetic length of the map was sensitive to errors. Using the uncorrected data, the genetic map showed a length of 1441 cM by the traditional approach but 1308 cM when constructing the map allowing for an error rate of 0.7%. Once the detected errors were corrected, the genetic length was 1299 cM by the traditional approach and 1274 cM when allowing for a residual error rate of 0.1%. Overall, the error detection method largely avoided the problem of map expansion and seems to have readily identified the great majority of errors.

DISCUSSION

Our error detection method is based on the simple approach of incorporating laboratory typing error into maximum likelihood map construction through the use of a penetrance function. For denser genetic maps, the method avoids the problem of map expansion and identifies potential errors so that they can be rechecked and corrected.

We are now using this approach routinely in the construction of dense genetic maps in the mouse and rat. As newly typed loci are added to our database, they are automatically analyzed to identify probable errors—allowing immediate rechecking. In practice, this procedure appears to reduce our error rates by nearly a factor of 10.

Some variations on the method can be easily envisaged: (1) The penetrance function could incorporate information about different error rates across loci (e.g., SSLPs may be harder to score than RFLPs) or across genotypes at a single locus (e.g., one homozygous genotype may be more likely to be incorrectly scored as another homozygous genotype than as a heterozygous genotype); (2) the assumed error rate ϵ could be treated as an unknown parameter to be estimated; (3) the analysis could be modified to incorporate the consequences of crossover interference. Crossover interference was neglected above for computational ease, but the existence of crossover interference makes apparent double crossovers even less likely than in our calculations above and thus increases the power to detect errors. Even without these modifications, however, the simple approach presented here seems to capture most of the power of the method.

Having demonstrated the value of the method for analyzing F2 intercrosses, a next challenge will be to apply it to the CEPH pedigrees used for human linkage analysis. From a theoretical standpoint, the generalization is straightforward. From a practical standpoint, however, it may pose a computational challenge: direct implementation of the method (i.e., simultaneously allowing every typing to be potentially erroneous) would require running time proportional to $m4^{2n}$ to compute the likelihood, where m is the number of loci and n is the number of offspring per family (Lander and Green, 1987). Possible solutions include (1) detecting errors separately in each child, while assuming that other children are correct (which would increase running time by no more than a factor of 16); (2) using an alternative algorithm to compute the exact likelihood, possibly to trim terms with low likelihoods; (3) detecting errors by estimating the posterior probability of error for each typing by using a Gibbs sampling approach (Geman and Geman, 1984) with our likelihood function; or (4) using a sufficiently fast computer, inasmuch as n is at most 15 in the CEPH families. Leaving aside the algorithmic issues, the statistical power to detect errors will depend on the mean spacing between *informative* markers in a typical meiosis.

In summary, laboratory typing error has posed minor problems in the construction of sparse genetic linkage maps. With the construction of ever denser maps, the situation changes dramatically, and it becomes important to devise automatic methods for detecting and correcting errors. The methods described here should assist in this goal, at least for crosses in experimental organisms.

ACKNOWLEDGMENTS

We are indebted to Bill Dietrich and Hillary Katz, who generated all the data for the mouse genetic linkage map used in the analysis above and to Mark J. Daly for work on the MAPMAKER computer package. This work was supported in part by grants from the National Institutes of Health (P50HG00098, R01HG00126, and R01HG00316 to E.S.L.), the National Science Foundation (DIR8611317 to E.S.L.), and the Markey Foundation (to E.S.L.).

REFERENCES

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragent length polymorphism. *Am. J. Hum. Genet.* **32:** 314–331.

Buetow, K. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *Am. J. Hum. Genet.* **49:** 985–994.

Chakravarti, A., and Lasher, L. K. (1992). Estimation of chromosome lengths under genotyping errors. (manuscript)..

Copeland, N. G., and Jenkins, N. A. (1991). Development and applications of a molecular genetic linkage map of the mouse genome. *Trends Genet.* **7:** 113–118.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. *Genetics* **131:** 423–447.

Donis-Keller *et al.* (1987). A genetic map of the human genome. *Cell* **51:** 319–337.

Dracopoli, N. C., O'Connell, P., Elsner, T. I., Lalouel, J.-M., White, R. L., Buetow, K. H., Nishimura, D. Y., *et al.* (1991). The CEPH consortium linkage map of human chromosome 1. *Genomics* **9:** 686–700.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and Bayesian restoration of images. *IEEE Pattern Anal. Mach. Intell.* **6:** 721–741.

Lander, E. S., and Green, P. (1987). Construction of multi-locus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84:** 2363–2367.

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., and Newburg, L. (1987). MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1:** 174–181.

Morton, N. E. (1991). Parameters of the human genome. *Proc. Natl. Acad. Sci. USA* **88:** 7474–7476.

Ott, J. (1977). Linkage analysis with misclassification at one locus. *Clin. Genet.* **12:** 119–124.

Ott, J. (1985). "Analysis of Human Genetic Linkage," Johns Hopkins University, Baltimore.

Patterson, D. (1991). Report of the Second International Workshop on Human Chromosome 21 mapping. *Cyotgenet. Cell Genet.* **57:** 168–174.