Am. J. Hum. Genet. 64:1739–1753, 1999

On the Assessment of Statistical Significance in Disease-Gene Discovery

Lue Ping Zhao,[1] Ross Prentice,[1] Fumin Shen,[2] and Li Hsu[1]

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle; and [2]Department of Epidemiology, Shanghai Medical University, Shanghai

Summary

One of the major challenges facing genome-scan studies to discover disease genes is the assessment of the genomewide significance. The assessment becomes particularly challenging if the scan involves a large number of markers collected from a relatively small number of meioses. Typically, this assessment has two objectives: to assess genomewide significance under the null hypothesis of no linkage and to evaluate true–positive and false–positive prediction error rates under alternative hypotheses. The distinction between these goals allows one to formulate the problem in the well-established paradigm of statistical hypothesis testing. Within this paradigm, we evaluate the traditional criterion of LOD score 3.0 and a recent suggestion of LOD score 3.6, using the Monte Carlo simulation method. The Monte Carlo experiments show that the type I error varies with the chromosome length, with the number of markers, and also with sample sizes. For a typical setup with 50 informative meioses on 50 markers uniformly distributed on a chromosome of average length (i.e., 150 cM), the use of LOD score 3.0 entails an estimated chromosomewide type I error rate of .00574, leading to a genomewide significance level >.05. In contrast, the corresponding type I error for LOD score 3.6 is .00191, giving a genomewide significance level of slightly <.05. However, with a larger sample size and a shorter chromosome, a LOD score between 3.0 and 3.6 may be preferred, on the basis of proximity to the targeted type I error. In terms of reliability, these two LOD-score criteria appear not to have appreciable differences. These simulation experiments also identified factors that influence power and reliability, shedding light on the design of genome-scan studies.

Received December 17, 1998; accepted for publication April 6, 1999; electronically published May 4, 1999.

Address for correspondence and reprints: Dr. Lue Ping Zhao, Quantitative Genetic Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1140 Fairview Avenue North, Seattle, WA 98105. E-mail: lzhao@fhcrc.org

© 1999 by The American Society of Human Genetics. All rights reserved.
0002-9297/99/6406-0028$02.00

Introduction

Advances in the development of high-throughput technologies, including chip technologies (Chee et al. 1996; Wang et al. 1998), promise that the scanning of human genomes in the study of complex disease traits will be routine in the future. A typical genome-scan study may involve 300–3,000 marker loci that cover the entire human genome, generating a large amount of marker data. Although a high marker density throughout the genome is expected to improve the power to discover disease genes, having a large number of markers on a small number of informative meioses also presents statistical challenges. One of them is the assessment of the genomewide significance level, an old problem in these modern times. On the basis of the theoretical consideration for sequential testing, Morton (1955) showed that LOD score 3.0 may be used as a cutoff for declaration of linkage in the mapping genetic traits by means of relatively few markers. Nowadays, however, with the use of densely distributed markers in the mapping of complex traits, the use of LOD score 3.0 may yield an excessive number of false-positive errors.

Recognizing this problem, Lander and Kruglyak (1995) proposed raising the cutoff from LOD score 3.0 to LOD score 3.6, leading to an ongoing debate. Witte et al. (1996) pointed out that LOD score 3.6 is arbitrary and that such a general guideline with a simple criterion may not serve the intended purpose. Curtis (1996), on the other hand, argued that this guideline is not useful, since it does not account for multiple models and multiple phenotypes tried in the analysis. Rao (1998) and Morton (1998) have been concerned that the proposed LOD-score-3.6 criterion would increase the chance that important linkage signals would be missed. Furthermore, via a whole-genome simulation study, Sawcer et al. (1997) show that the use of LOD score 3.2 ensures a genomewide significance level of 5%, leading them to argue in support of the LOD-score-3.0 criterion.

It appears that both the type I error and false-positive prediction error (FPP), which are related yet different indices, have been used in this debate. As a matter of definition, the type I error is the error of rejecting the null hypothesis, on the basis of a specific test statistic, *under the null hypothesis*. FPP, on the other hand, is the

error of incorrectly predicting the genomic regions that include a disease gene locus, after the null hypothesis has been rejected. For example, Sawcer et al. (1997) evaluated the type I error on the basis of the maximum LOD score, at the chromosome level, under the null hypothesis that the chromosome does not contain any disease genes. They then combined all chromosomewide significance levels to compute the genomewide significance, under the null hypothesis that the genome does not contain any disease gene. On the other hand, Morton (1998) emphasized the importance of computing FPP, along with the true-positive prediction rate (TPP), and also suggested that the TPP/FPP ratio, the basis for calculation of the reliability index, is essential for choosing the criterion. Furthermore, Lander and Kruglyak (1995) computed the expected number of false-positive errors under the null hypothesis and used it as a measure of FPP under alternative hypotheses.

Separate consideration of type I error versus FPP may be a step in the right direction toward resolution of the debate about the desirable LOD-score value. In the remainder of this report, we elaborate on definitions of the type I error, FPP, and expected number of false-positive signals, in addition to several other indices, and discuss their relationship in the assessment of genomewide significance. Following those definitions, we describe the genomewide, chromosomewide, and pointwise significance levels and discuss their relationships. Recognizing the small sample sizes used in typical genome-scan studies, we choose the Monte Carlo simulation method to evaluate relevant statistics. Via these Monte Carlo experiments, we evaluate LOD scores 3.0 and 3.6 in terms of the type I error, power, reliability, TPP, and FPP. Some discussion and conclusions follow.

Type I Error, Expected Number of False-Positive Errors, FPP, and Related Indices

Three commonly used indices for the assessment of genomewide significance are type I error, expected number of false-positive linkage signals, and FPP. For clarity, we will define these three indices here. The genomewide type I error is defined as the probability of falsely rejecting the null hypothesis that the genome does not include any disease gene, when this null hypothesis is actually true. Specifically, consider a null hypothesis (H_0) that the genome does not include any disease gene, as presented in the second column of table 1. When the test statistic for measurement of the genomewide significance level exceeds a threshold value, the statistical decision is to reject the null hypothesis. This rejection is, of course, an error, which could mislead investigators to search for nonexistent disease genes. Formally, the LOD-score type I error may be written as

Table 1

Definition of Types I and II Errors

	Error Under	
	H_0	H_1
H_0 not rejected	$1-\alpha$	β (type II error)
H_0 rejected	α (type I error)	$1-\beta$ (power)
Total	1	1

$$\Pr[\max_{\text{genomewide}} (\text{LOD scores}) > T_\alpha | H_o] = \alpha ,$$

where the maximization is over all markers genomewide, α is the designated type I error rate (e.g., $\alpha = .05$), and T_α is the corresponding cutoff value for the maximum LOD scores (e.g., $T_{.05} \cong 3.6$).

Besides deciding whether to reject the null hypothesis, one can also count how many times linkage-test statistics exceed the designated threshold (e.g., LOD score 3.6). The expectation of this count is defined as the expected number of false-positive linkage signals. It has been shown that such a count, provided that there are a small expected number of occurrences, approximately follows a Poisson process (Feingold et al. 1993). The derivation of this approximation requires several assumptions. One key assumption concerns dependencies among recombination indicators, which are influenced both by physical distances between markers and by interference between genomic regions. Owing to the complexity of the human genome, modeling this dependence is difficult, and yet this dependence determines how often linkage-test statistics in adjacent marker loci simultaneously exceed the threshold value. Furthermore, an increase in the number of markers, increases dependencies of recombination events in adjacent marker intervals, hence leading to an increased number of false-positive linkage signals, as will be illustrated in our simulation study. In light of both concerns, one may prefer the type I error to the expected number of false-positive errors, as a measure of genomewide significance.

The key objective of the genome-scan study is not only to test the null hypothesis but also to predict which genomic regions include disease genes, after the null hypothesis is rejected. For example, after rejecting the null hypothesis, one finds that the test statistic, at the kth locus, exceeds the threshold value (e.g., LOD score 3.6) and thus predicts that the genomic interval, $(\delta_k - \varepsilon, \delta_k + \varepsilon)$, includes a disease gene, where δ_k denotes the map distance of the kth marker on the human genome map and a constant ε is used to define the targeted genomic region (e.g., 10 cM). It is important to realize that the choice of this constant ε is for the convenience of this simulation study and is somewhat arbitrary, since additional information may be used in practical genome-

scan studies (see the Discussion section). Of course, such a prediction could prove to be either true or false, as illustrated in table 2, in which f and t are used to denote the false and true prediction rates, respectively, and the subscripts p and n denote positive prediction and negative prediction, respectively. Note that t_p is TPP and that f_p is FPP. Formally, FPP may be expressed as

$$\text{FPP} = \Pr[\text{no gene in } (\delta_k - \varepsilon, \delta_k + \varepsilon) \text{ at every locus having } T_k > T_\alpha \mid H_1] ,$$

where T_k is the LOD score at the kth locus and H_1 is a specific alternative hypothesis (e.g., existence of one or more disease genes on the same chromosome). A similar definition may be given to TPP. Although conceptually different, type I error and FPP are intrinsically connected. Specifically, when a specific alternative hypothesis is close to the null hypothesis, the summation of TPP and FPP is approximately equal to the type I error. Otherwise, they measure different aspects; one measures errors in testing, and the other measures errors in prediction.

An alternative to the use of FPP is to measure the probability of making wrong predictions only among those predictions. Formally, this alternative measure may be defined via a conditional probability; that is,

$$\Pr[\text{no gene in } (\delta_k - \varepsilon, \delta_k + \varepsilon) \text{ at every locus having } T_k > T_\alpha \mid \max(T_k) > T_\alpha, H_1] .$$

This measurement is obtainable as the ratio of FPP versus the power of rejecting the null hypothesis, to be defined in the following discussion.

Closely related to measurements of significance level are those for power in the testing of the hypothesis and for TPP in the prediction of candidate genomic regions. With regard to hypothesis testing, the power is used to measure the probability of rejecting the null hypothesis under a specified alternative hypothesis, presented in the third column of table 1. Formally, the power may be expressed as

$$\text{Power} = \Pr[\max_{\text{genomewide}} (\text{LOD scores}) > T_\alpha | H_1] ;$$

that is, the probability that the maximum LOD score exceeds the cutoff value under the alternative hypothesis. The power is generally expressed as $1 - \beta$, where β is used to denote the type II error of not rejecting the null hypothesis under the alternative hypothesis. By definition, the power is dependent on the alternative hypothesis.

With regard to prediction, an important measurement is TPP, t_p, as described above. Morton (1998) suggests that the ratio $t_p/(t_p + f_p)$ or t_p/f_p, as a reliability measure, is a more important index than either t_p or f_p alone, in the evaluation of a procedure for discovery of genes. In the remainder of this report, we will use $t_p/(t_p + f_p)$ as a reliability measure. Interestingly, both ratios, $f_p/(t_p + f_p)$ and $f_n/(t_n + f_n)$, have been used as measures of prediction errors in the screening literature (e.g., Rothman and Greenland 1998).

Table 2

Definition of TPP and FPP, with True-Negative Prediction Rate and False-Negative Prediction Rate

	False Prediction	True Prediction	Total
Negative prediction	f_n	t_n	$f_n + t_n$
Positive prediction	f_p (FPP)	t_p (TPP)	$f_p + t_p$
Total	$f_n + f_p$	$t_n + t_p$	1

Genomewide, Pointwise, and Chromosomewide Significance

Genomewide versus Pointwise Significance

Pointwise significance is defined as the probability that, just by chance, one would encounter a test statistic as large as or larger than that observed at a specific locus, under the null hypothesis of no linkage at that locus, whereas genomewide significance is defined as the probability that one would encounter such an extreme test statistic somewhere in a whole-genome scan, under the null hypothesis of no disease gene genomewide (Lander and Kruglyak 1995). Under the null hypothesis, the genomewide significance level could be assessed on the basis of the pointwise significance levels and the Bonferroni correction. However, markers on the same chromosome are highly correlated, given the chromosomal architecture, so that the Bonferroni procedure may be unacceptably conservative. Furthermore, the correlation of recombination events between markers is influenced by many factors, including hot spots and interference. The complexity of the correlation structure generally makes it difficult to accurately derive a joint distribution of all pointwise test statistics from the same chromosome. Without a joint distribution, it is not possible to analytically determine chromosomewide significance on the basis of pointwise significance.

Genomewide versus Chromosomewide Significance

Cytogenetically, a human genome consists of 22 pairs of autosomes and a pair of sex chromosomes. In meiosis, all these chromosomes are separately transmitted from parents to their children. Statistically, such a biological

fact implies that events on distinct chromosomes are independent from each other; hence, the genomewide significance level is determined by chromosomewide significance level. Specifically, genomewide significance is the probability that one will observe a test statistic more extreme than that observed under the null hypothesis that the genome does not contain any disease-gene loci; chromosomewide significance is the counterpart at the level of each chromosome. Let $P_{\text{chromosome } i}$ denote significance at the ith chromosome, one of 24 distinct chromosomes (i.e., 22 autosomes, 1 X chromosome, and 1 Y chromosome). A simple probability argument shows that the genomewide significance, $P_{\text{genomewide}}$, may be obtained as follows (Sawcer et al. 1997):

$$P_{\text{genomewide}} = 1 - \prod_i (1 - P_{\text{chromosome } i}) \ . \qquad (1)$$

Given the significance level on each chromosome, equation (1) allows one to compute the genomewide significance level. Conversely, given the critical value for the genomewide significance level, one can compute the critical value for chromosomewide significance level, provided that a common error rate is specified for each chromosome; for example, suppose that the genomewide type I error rate is considered to be .05 and that a common type I error rate is chosen for each chromosome. Equation (1) then simplifies to $1 - (1 - P_{\text{chromosome}})^{24} = .05$, and solving this equation leads to a chromosomewide significance of ~.0021, which is comparable to the Bonferroni correction (≅.05/24). Alternatively, one may allow the type I error rate on each chromosome to be proportional either to its length or to the density of markers covering individual chromosomes. Under such scenarios, one may use variable type I error rates depending on the chromosome of interest, while retaining a specific genomewide type I error rate.

Also because of the independence between chromosomes, it suffices to consider the genomewide significance level via consideration of the chromosomewide significance level. Hence, in the remaining discussion, we will focus on only the chromosomewide significance, with attention to the variation in chromosome length. For simplicity, we choose to assign the type I error rate of .0021 to each chromosome.

The Simulation Setup for Monte Carlo Experiments

The essence of discovering disease genes via linkage analysis is to estimate recombination fractions, which measure distances between a disease-gene locus and linearly ordered multiple genetic markers whose relative locations are known. In human studies, the estimation procedure is complicated, and related computations are extensive, because meiotic events are not directly observable. To circumvent this computational challenge, we consider the situation in which all meioses are fully informative. For each informative meiosis with K markers, with or without one or more disease-gene loci, one can simulate recombination events between adjacent marker loci, according to a Bernoulli process having a mean that is equal to the recombination fraction. These recombination events are generally assumed to be independent, corresponding to the absence of interference. To generate a recombination-indicator variable between a disease locus and every marker, one thus counts how many recombination events occur between them; an odd number of recombination events indicates the presence of a recombination event, and an even number indicates its absence. To allow for unlinked disease loci due to genetic heterogeneity or epistatic interactions, we place a disease locus on the far left or far right and simulate a recombination event with the adjacent marker locus, using recombination fraction .50. Hence, such a disease locus is not linked with any of the marker loci under consideration. The actual available data from the simulation experiment are the mixture of these different types of recombinant data, with the mixture proportion being λ. The proportion λ has a value of 0–1, with 0 representing the situation corresponding to the null hypothesis and with 1 representing the situation in which a disease-related gene that is pertinent to the entire study population. With essentially the same setup, the simulation experiment was also extended to include a second disease gene, with an additional λ value.

The presence of interference implies that recombination events across the genome may no longer be independent and that the dependence structure could be rather complicated. To study the influence that interference has on the cutoff, we consider a simplistic interference model. First, independent recombination events are simulated on the genome, as has been described above. Then simulated recombination events are modified to yield the final realization of the recombination events: if, initially, there is no recombination event in a particular interval, the final realization for the interval has a probability of being a recombinant, provided that an adjacent interval has a recombination event. Now the probability is treated as a dependence parameter, with 0 representing the absence of interference and with values of >0 for representing the presence of interference.

Besides the dependence parameter and the mixing proportion, there are several other key parameters. Since the experiment simulates markers from a single chromosome, it is natural for us to consider a chromosome of average length,ι = 150 cM, whereas the length of a specific chromosome may be 60–300 cM. Another parameter is the number of markers (η) used in the genome scan, a value that, in most simulations, we chose to be η = 50 markers/chromosome, with a range of η = 20–100/chromosome and with a range of 7.5–1.5 cM rep-

resenting the corresponding average distance between markers on a 150-cM chromosome. Such densities imply that 480–2,400 markers are used in the genome-scan study. Furthermore, the number of informative meioses, n, reflects the sample size of the study and typically is considered to be $n = 50$, with a range of $\eta = 30$–150. Finally, in predicting the candidate genomic region, one needs to assign a neighborhood around the kth linked marker locus at the position, δ_k; that is, the candidate genomic region is predicted as being $(\delta_k - \varepsilon, \delta_k + \varepsilon)$. Obviously, the narrower the region, the more precise the prediction, at the expense of additional false–positive errors. In a typical simulation, $\varepsilon = 10$ cM, with a range of 2–20 cM.

Consider a genome-scan study with K marker loci on n informative meioses, in which the ith vector of simulated recombination events may be denoted as follows: $y_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$. Averaging these n vectors results in a point estimate of recombination fractions of a putative disease gene with all K marker loci. The pattern of these estimated recombination fractions offers a clue to the position of putative disease genes. Figure 1 shows examples of patterns of estimated recombination fractions and LOD scores resulting, in one case, from the presence of one major gene located at ~40 cM (fig. 1, left-side panels) and, in another case, from the presence of two major genes, one at ~40 cM and the other at ~120 cM (fig. 1, right-side panels). Clearly, under the alternative with either one major gene or two major genes, recombination fractions in the neighborhood of major genes are <.50, as expected. Interestingly, it appears that marker loci, in the same region as but distant from those major genes, often have recombination fractions that are <.50. This phenomenon is referred to as a "region effect." Furthermore, in the case of two major genes, the patterns both of recombination fractions and of LOD scores appear to suggest that there is one disease gene and that it is located at 50–100 cM—which is, of course, incorrect. Visual examination of genome-scan data can help one to identify certain problems and issues in the interpretation of genome-scan studies. As a service to interested investigators, we have set up a program at the Fred Hutchinson Cancer Research Center QGE Website), with which readers may simulate genome-scan data and gain further experience with potential patterns of recombination fractions and LOD scores. Besides having interesting patterns, recombination events among adjacent marker loci are highly correlated, because of the genomic structure underlying the K recombination events. Correlation between markers decays as distances between them increase, and the pattern with one major gene is shown in figure 2, in which the average density among 50 markers is 3 cM and in which the color intensity corresponds to the magnitude of correlation. Although the Markovian correlation model may be used

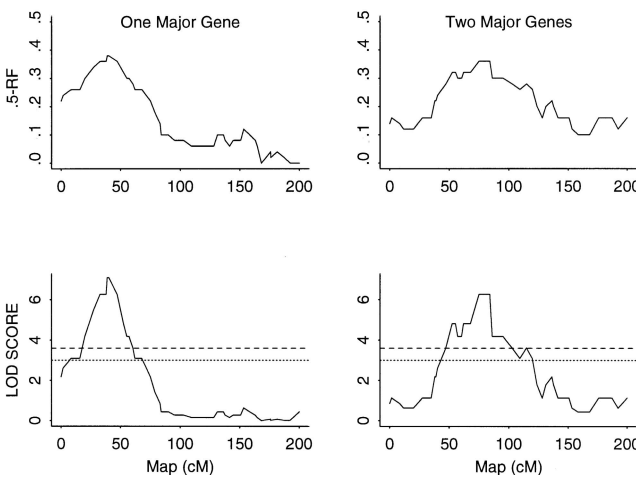

Figure 1 Typical pattern of estimated recombination fractions from simulated genome scan with the mixing proportions $\lambda = 0$, .25, .50 .75, and 1.00 (other examples may be simulated online at the Fred Hutchinson Cancer Research Center QGE Website).

to approximate such a correlation structure (e.g., see Feingold et al. 1993), this approximation is appropriate only if the number of informative meioses is large and only if there is no interference throughout the genome. Similarly, the same conditions are required for approximation, by Brownian motion, of the random process of y_i.

Because of the small sample sizes and the complex structure of the recombination process, it appears that the only available option in the study of the statistical-significance issue is the Monte Carlo simulation method. The Monte Carlo experiment used here involves 100,000 replicates, so that the $\alpha = .0021$ estimation of the type I error is based on 210 (i.e., 100,000 × .0021) expected events, whereas the estimated standard error of the type I error rate 1.45×10^{-4} (i.e., $\approx \sqrt{.0021 \times (1 - .0021)/100{,}000}$).

Evaluation of LOD Scores 3.0 and 3.6

Using the simulation setup described above, we evaluate two proposed criteria—namely, LOD 3.0 and 3.6—for both testing the hypothesis and predicting the location of disease genes. The evaluation includes comparisons of type I errors, powers, reliabilities, TPP, and FPP.

The Testing-and-Prediction Procedure

For the purpose of testing the null hypothesis, we compute LOD scores at all marker loci, which is equivalent to a two-point linkage analysis for scanning the genome. When the maximum LOD scores exceed a threshold (e.g., 3.0 or 3.6, which are used in the present study),

the null hypothesis is rejected—that is, there seems to be one or more disease-related genes on the chromosome. Proceeding to the prediction stage, one could look for marker loci that have LOD scores exceeding the same threshold value. As expected, multiple marker loci may be candidates, either (*a*) because they are closely linked with the underlying disease gene and because recombination events among markers within a neighborhood are correlated, thereby inducing a "regional effect," or (*b*) because they occur purely by chance. To balance between specificity and FPP while avoiding the influence of local correlation, we adopt an iterative procedure: (1) we identify the genomic region around the marker locus with the maximum LOD score—that is, $(\delta_k - \varepsilon, \delta_k + \varepsilon)$; then (2), after excluding the identified region, we search for the maximum of LOD scores on remaining markers, and, if the maximum LOD score exceeds the threshold value, consider the corresponding marker locus and its neighborhood region; and (3) we continue this process until no LOD score exceeds the threshold value.

Type I Error under the Null Hypothesis

Under the null hypothesis, the simulation study has shown that, for a typical simulation setup—that is, 50 informative meioses, 50 markers, and an average chromosome size of 150 cM—the chromosome-specific type I errors are .00574 and .00191, corresponding to LOD scores 3.0 and 3.6, respectively (table 3). The former value is >.0021, implying that LOD score 3.0 is liberal; the latter value, on the other hand, is less than the designed type I error, indicating that LOD score 3.6 is slightly conservative.

The Monte Carlo experiments show that, for n = 30–100, the type I error corresponding to LOD score

Table 3

Estimated Type I Errors under a Typical Simulation Setup and Its Variations, with Number of Informative Meioses, Length of Chromosome, and Number of Markers

	Estimated Type I Error, for	
	Lod Score 3.6	Lod Score 3.0
Typical setup[a]	.00191	.00574
No. of informative meioses:		
30	.00117	.00566
50	.00191	.00574
60	.00066	.00247
100	.00081	.00369
Chromosome length:		
60 cM	.00098	.00292
150 cM	.00191	.00574
200 cM	.00200	.00680
300 cM	.00263	.00856
No. of markers:		
20	.00132	.00422
40	.00190	.00597
50	.00191	.00574
100	.00224	.00706

[a] 50 meioses, 50 markers, 150-cM chromosome.

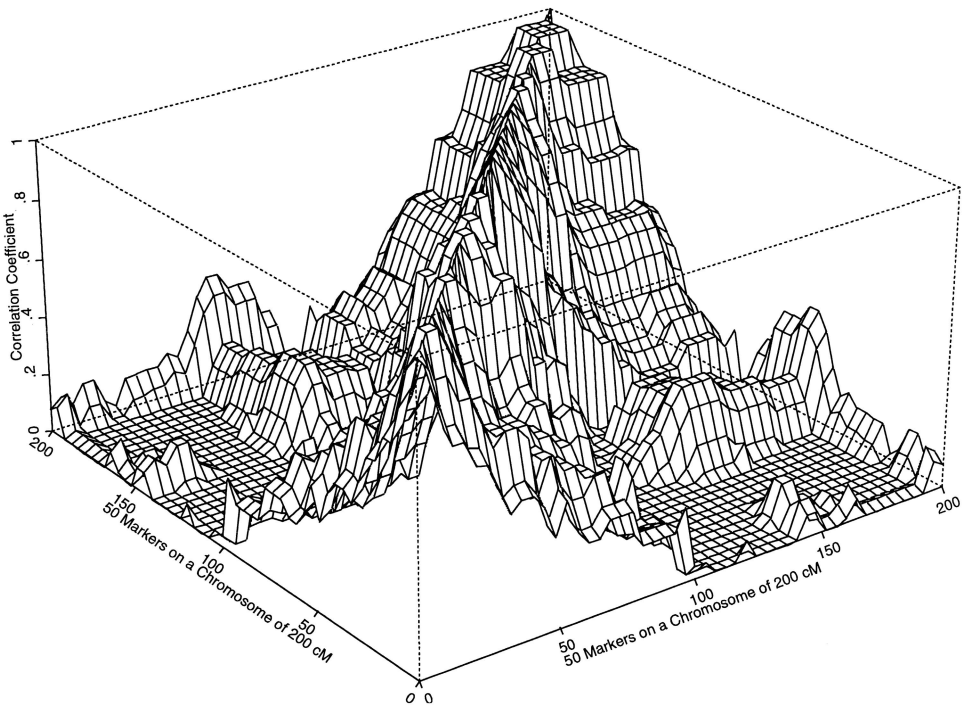

Figure 2 Typical pattern of estimated correlation coefficients from a simulated genome scan with mixing proportion λ = .75

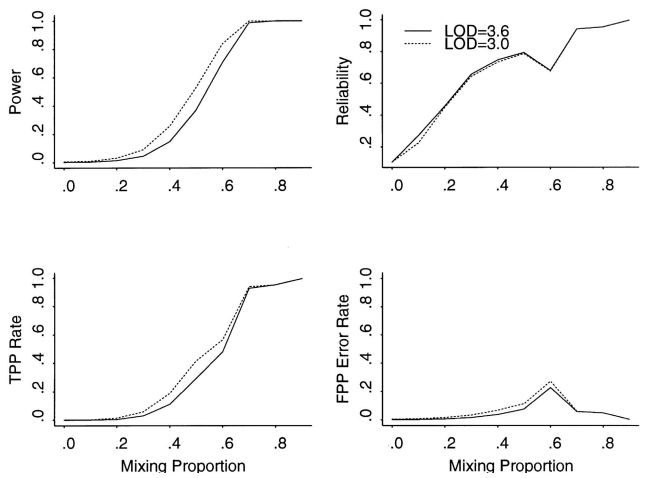

Figure 3 Powers, reliability, TPP, and FPP, for LOD scores of 3.0 and 3.6 in the presence of one major gene.

3.6 is .0017–.00081, whereas the type I error corresponding to LOD score 3.0 ranges is .00566–.00369 (table 3). Similarly, the type I error is smaller on shorter chromosomes than on longer chromosomes. On the other hand, it is comforting to know that an increase in the number of markers has a fairly limited impact on the type I error. Our results are consistent with the findings by Lander and Kruglyak (1995), as well as with the empirical observation made by Sawcer et al. (1997); in fact, it also provides an explanation for the latter—that is, after account is taken of all the different chromosomal lengths and of the actual sample size, the desired LOD score in the Sawcer et al. (1997) study is ~3.2, or somewhere between LOD sore 3.0 and LOD score 3.6.

Comparison under an Alternative Hypothesis, with One Major Gene

Now we will study these two critical values under the alternative hypothesis with one major gene. Figure 3 has four panels: the upper-left panel describes the estimated power of rejecting the null hypothesis, the upper-right panel shows estimated reliability, the lower-left panel shows empirical TPP, and the lower-left panel shows estimated FPP. The unbroken line represents the results for the LOD-score-3.6 criterion, whereas the dotted line represents the values for the LOD-score-3.0 criterion. Comparing the two power curves, one observes the expected gain of power when the LOD-score-3.0 criterion, rather than the LOD-score-3.6 criterion, is used (fig. 3, upper-left panel). After rejecting the null hypothesis, one is interested in either TPP and FPP or, simply, the reliability index, per Morton's (1998) suggestion. The upper-right panel of figure 3 appears to suggest that reliabilities for LOD scores 3.6 and 3.0 are comparable and that the former appears to be marginally favored. On investigating TPP (fig. 3, lower-left panel), one would see that TPP increases as the mixing proportion λ increases. Overall, TPP is higher for LOD score 3.0 than for LOD score 3.6, as expected. Interestingly, FPP increases with the mixing proportion until λ is ~.6 and then decreases. One possible explanation is that, as the mixing proportion increases to .6, FPP increases because of the "regional effects" described above and that yet, after the mixing proportion becomes >.6 , the ratio of linkage signal versus regional effects increases and thus reduces FPP. Because of a close connection between the regional effects and the width of the prediction intervals from $-\epsilon$ to $+\epsilon$, choosing ϵ on the basis of patterns of either LOD scores or recombination fractions may offer additional information and hence may reduce FPP (see the Discussion section). Addressing this issue is important, since FPP here could be as high as 20%. In comparison, FPPs associated with LOD score 3.0 are higher than those with LOD score 3.6. It is recognized that, under the alternative hypothesis, properties may depend on the single major gene's location being specified as 40 cM. To broaden the alternative hypothesis, a separate simulation experiment has been performed with a location chosen as a uniform variate over the length of the chromosome in each replicate (not shown). The resulting powers, reliabilities, TPP, and FPP are comparable to those observed for the fixed location.

Comparison under an Alternative Hypothesis, with Two Major Genes

In disease-gene discovery, one needs to allow two or more major genes on the same chromosomes, as alternative hypotheses. For simplicity, we consider alterna-

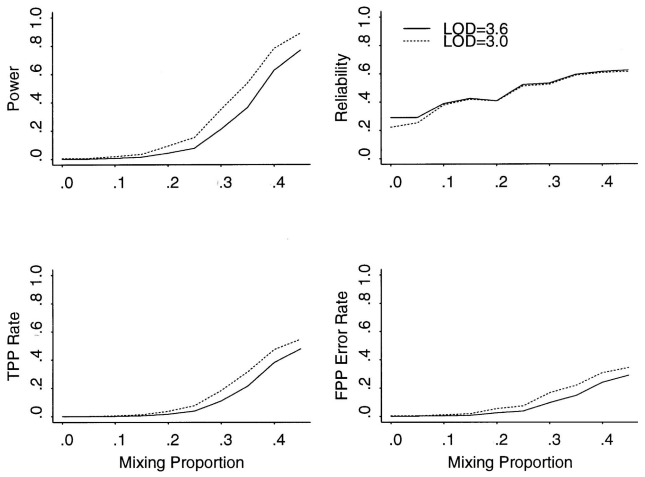

Figure 4 Powers, reliability, TPP, and FPP, for LOD scores of 3.0 and 3.6 in the presence of two major genes.

tives with two major genes, induced by mixing three different recombination events: (1) the presence of a major gene at 40 cM, (2) the presence of the second major gene at 120 cM, and (3) the absence of any major genes. We use two equal mixing proportions; for example, $\lambda = .3$ implies that 30% of recombination is linked with the first gene, 30% with the second gene, and the remaining 40% with no gene at all. Figure 4 shows the power, reliability, TPP, and FPP. As expected, the power of testing the null hypothesis is lower for the two-major-genes alternative than that for testing a single major gene, and use of LOD score 3.0 provides higher power than does use of LOD score 3.6 (fig. 4, upper-left panel). The reduced power, compared with that for a single disease gene, is probably due to the presence of fewer informative meioses at a particular major gene, in contrast with what is seen for the single-major-gene alternative. Although not entirely surprising, the reliability for LOD scores 3.0 and 3.6 is comparable and is lower than that under the single-major-gene alternative (fig. 4, upper-right panel). On further investigation of TPP and FPP (fig. 4, lower panels), it becomes clear that the low reliability is caused by a low TPP in combination with a high FPP. This result suggests that, even after detection of a linkage signal, it is difficult to localize those disease genes when there are two (or more) disease genes on the same chromosome. As noted above, the locations of the two major genes under the alternative hypothesis are specified as being at 40 and 120 cM. To broaden this alternative hypothesis, we perform a separate simulation experiment with uniformly distributed locations for the paired disease gene in each replicate (data not shown). Although comparable to those for the fixed locations, the resulting powers, reliabilities, TPP, and FPP appear to have greater variability than those for the fixed locations, probably because of the variable distances between the two disease genes.

Influences of ϵ under the One-Major-Gene Alternative

As noted above, the patterns of the power curves for LOD score 3.0 are comparable to those for LOD score 3.6. For simplicity, the evaluation of power hereafter concentrates on LOD score 3.6. As described above, the prediction procedure is based on LOD scores; if the maximum LOD score is >3.6, then one predicts that the region around the corresponding marker, $(\delta_m - \varepsilon, \delta_m + \varepsilon)$, includes a disease locus. This prediction is correct if the candidate region does include the true disease-gene locus and is incorrect otherwise. In the interest of space, the following presentation focuses on power and reliability, when ϵ is chosen to be 2, 5, 10 or 20 cM, and the corresponding figures are shown in the upper and lower panels of figure 5*a*. As expected, the power of rejecting the null hypothesis is not influenced by the choice of ϵ, since the test statistic does not depend on this parameter (fig. 5*a,* upper panel). The reliability, on the other hand, favors a larger ϵ, because the prediction with a larger ϵ is less specific and, hence, more reliable (fig. 5*a,* lower panel). Interestingly, the reliability for the choice of $\varepsilon = 10$ is substantially better than that for the choice of $\varepsilon = 2$ or 5 cM and is not much worse than that for the choice $\varepsilon = 20$ cM. On further investigation of TPP, it appears that TPP for $\varepsilon = 10$ cM is quite close to that for $\varepsilon = 20$ cM and is substantially better than that for $\varepsilon = 5$ cM or 2 cM, again confirming the earlier observation based on the reliability (not shown).

Influences of Chromosome Length under the One-Major-Gene Alternative

Since human chromosomes have variable lengths, it is important to evaluate the influence that their variability has on the properties of the testing-and-prediction procedure. Here we consider four different chromosome lengths: 60, 150, 200, and 300 cM (fig. 5*b*). Since, throughout this simulation setting, the number of markers is fixed at 50, a variation of chromosome length essentially alters the correlation structure among recombination events; a shorter length implies a greater correlation. It appears that the test statistic's power has limited variation for various chromosome lengths, implying that the testing procedure is rather robust across all chromosomes (fig. 5*a,* upper panel). However, the reliability for longer chromosomes—for example, those that are ~300 cM—is rather variable and also is lower than that for chromosomes that are <200 cM (fig. 5*a,* lower panel). Meanwhile, reliability for those chromosome that are 60–200 cM is rather comparable, showing that the shorter chromosomes—that is, 20 of the 24 chromosomes—have comparable reliability.

Influences of Marker Numbers under the One-Major-Gene Alternative

In the discover of disease genes, one way to improve power is to increase the number of markers. For a chromosome of average length, an increase in the number of markers certainly increases the information content, but the increase in information content is not necessarily proportional to the increase in the number of markers, since the number of markers is correlated with an increase in the correlation of recombination events among marker loci. To address this issue, we consider four scenarios, with 20, 40, 50, and 100 markers used in genome scan, corresponding, respectively, to densities of, on average, 7.5, 3.7, 3.0, and 1.5 cM. Figure 5*c* shows power and reliability. Power as shown in the upper panel of figure 5*c* shows that an increase in the number of markers improves the power but that, surprisingly, the improvement is modest. However, from the perspective of

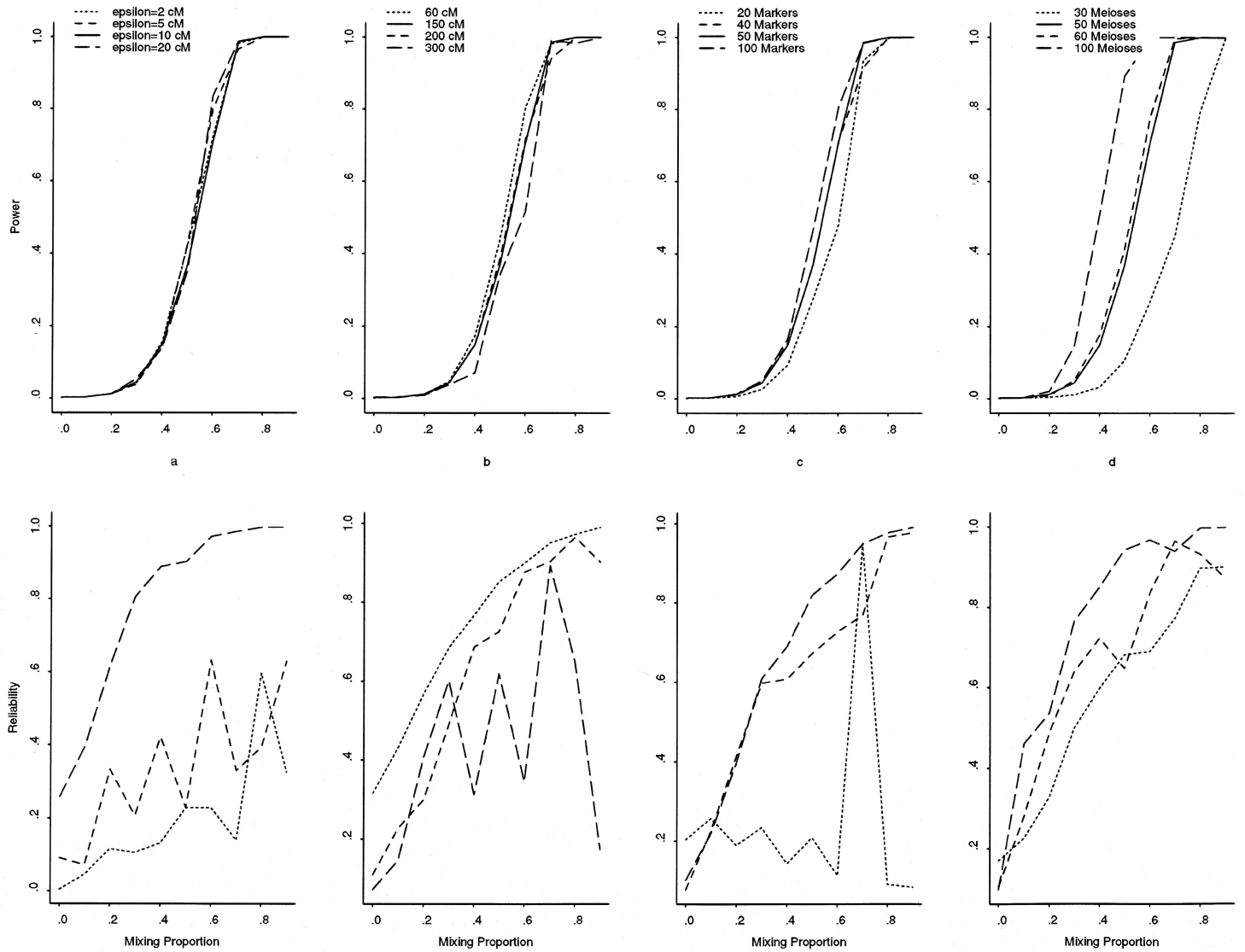

Figure 5 Powers and reliability for LOD score of 3.6 under the one-major-gene alternative, with four key parameters varying: (*a*) the neighborhood parameter for the candidate genomic region is chosen to be $\varepsilon = 2$, 5, 10, or 20 cM; (*b*) the length of the chromosome is 60, 150, 200, or 300 cM; (*c*) $\eta = 20$, 40, 50, or 100; (*d*) $n = 30$, 50, 60, or 100.

prediction, it appears that the presence of >20 markers improves the reliability substantially (fig. 5*c,* lower panel). On the other hand, there appears to be limited gain in reliability when there are >~40 markers.

Influences of Sample Sizes under the One-Major-Gene Alternative

Also for the purpose of improvement in power, one alternative to having more markers is to increase sample sizes—that is, having more family members and/or more families and thus increasing the numbers of informative meioses. To evaluate the gain in power, we consider four scenarios: 30, 50, 60, and 100 informative meioses. Fig. 5*b* shows powers and other related statistics. As expected, an increase in the number of informative meioses directly increases the powers of test statistics (fig. 5*c,* upper panel), the improvement of which is remarkable. The reliability is also improved by an increase in sample size (fig. 5*c,* lower panel), but the improvement is relatively modest. On further investigation of TPP and FPP (not shown), it becomes clear that the improvement of TPP when sample size increases is substantial, whereas FPP also appears to increase somewhat. Nevertheless, an increase in sample size is effective for improvement of the power to detect linkage signals and to predict where disease genes are, if the study of either more family members or more families is feasible and if no genetic heterogeneity is introduced.

Influences of Sample Sizes under the Two-Major-Genes Alternative

The simulation result shown above indicates that power and reliability become fairly low in the presence of two major genes. One possible cause for this reduction is that the number of informative meioses for each gene is reduced. To examine this issue, we consider four scenarios with *n:* $n = 50$, 75, 100, and 150. As expected, an increase in the sample size improves the power of rejecting the null hypothesis (fig. 6*a,* upper panel). Examination of the reliability figure (fig. 6*a,* lower panel) suggests that an increase in sample size also improves the reliability, but the magnitude of the improvement is relatively modest. Further examinations of TPP and FPP (not shown) suggest that an increase in sample size improves TPP substantially and, at the same time, increases FPP, diminishing the improvement of the reliability. Therefore, when two or more major genes are present on one chromosome, an increase in sample size provides power for rejection of the null hypothesis and improves the chance that disease genes will be detected, at the expense of having an increase in false-positive linkage signals.

Influences of the Number of Markers under the Two-Major-Genes Alternative

Again, on the basis of the same rationale as has been given above, one may increase the number of markers in the hope of increasing the chance to detect linkage signals. Paralleling the earlier specification, we consider four different scenarios: 20, 40, 50, and 100 markers used in scanning each chromosome of the genome. Consistent with the earlier finding under the one-major-gene alternative, the addition of markers improves the power (fig. 6*b,* upper panel), and the improvement becomes a marginal return once there are ⩾40 markers. Similarly, having more markers improves the reliability somewhat, with diminishing returns when $\eta > 20$ (fig. 6*b,* lower panel). Interestingly, the reliability seems quite variable with a small number of markers (e.g., 20).

Influences of Chromosome Length under the Two-Major-Genes Alternative

Recognizing the variation in chromosome length, we consider four scenarios for the latter: 60, 150, 200, and 300 cM and assess their influences on power and reliability (fig. 6*c*). Note that the second gene locates at 120 cM, outside the shortest chromosome, which is at 60 cM, and thus this scenario implies that the single gene is present at 40 cM. The power for chromosomes ⩾150 cM seems to be comparable (fig. 6*c,*upper panel). As expected, the power for shorter chromosomes is much lower, because it includes only a single gene with the much smaller mixing proportion λ. Similarly, the reliability for chromosomes ⩾150 cM is also comparable (fig. 6*c,* lower panel). This observation is consistently supported by TPP and FPP (data not shown). The exceptional reliability for shorter chromosome is due to extremely low FPP, which results in the presence of only one major gene.

Influences of ϵ under the Two-Major-Genes Alternative

A typical example of the two-major-genes alternative, shown in figure 1, suggests that the presence of two or more genes could easily create a chromosomewide pattern, of either recombination fractions or LOD scores, that would mislead the localization of those disease genes. One way to improve TPP naturally is to use less specific criteria in the identification of the candidate genomic region—that is, an $\epsilon > 10$ cM. To address the influences of ϵ, we consider the same four scenarios as have been given above: $\varepsilon = 2$, 5, 10, and 20 cM. As expected, the choice of ϵ does not affect power (fig. 6*d,* upper panel). Examining the reliability (fig. 6*d,* lower panel), one observes an appreciable improvement when a large ϵ is used. This improvement results from the combination of increased TPP and decreased FPP, when

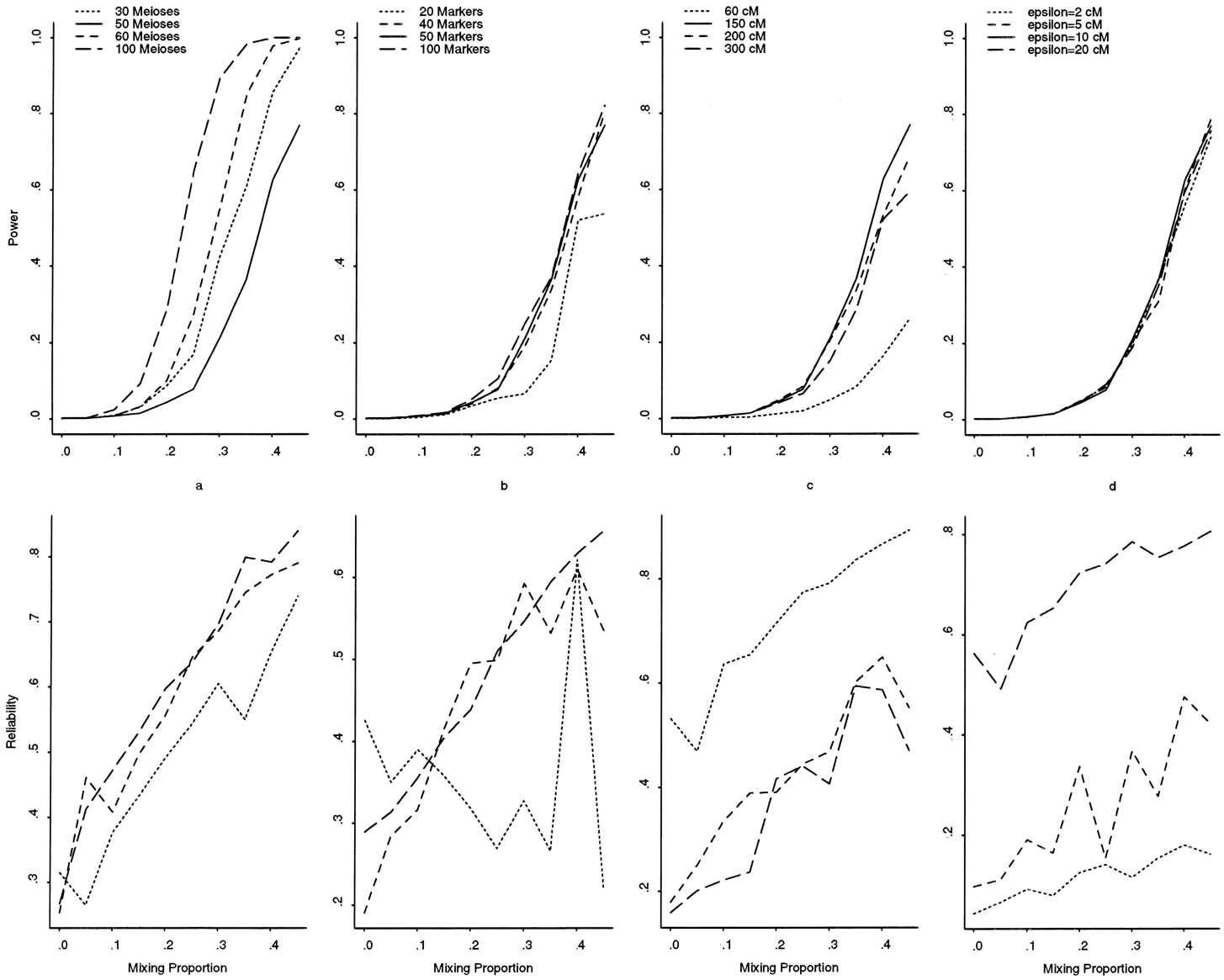

Figure 6 Powers and reliability for LOD score 3.6 in the presence of two major genes, with four key parameters varying: (*a*) $n = 30$, 50, 60, or 100; (*b*) $\eta = 20$, 40, 50, or 100; (*c*) length of chromosome is 60, 150, 200, or 300 cM; (*d*) neighborhood parameter for candidate genomic region is chosen to be $\varepsilon = 2$, 5, 10, or 20 cM.

ϵ increases (not shown). Increasing ϵ to >20 cM could further improve the reliability (not shown); however, such an improvement is, of course, at the expense of prediction, which becomes less precise.

Expected Number of False-Positive Errors and Number of Markers

As discussed above, in the section "Type 1 Error, Expected Number of False-Positive Errors, FPP, and Related Indices," the expected number of false-positive errors provides another sensible index for measurement of the error of incorrect rejection of the null hypothesis. The key assumption for its validity is that the number of error occurrences is "relatively" small (Lander and Kruglyak 1995). Here we consider a realistic situation that may violate this assumption; that is, when the number of markers in genome-scan studies increases, the number of error occurrences also increases, because of the correlation of recombinant events among adjacent marker intervals. Under the aforementioned simulation setup, we compute the average number of false-positive errors under the null hypothesis, using LOD scores 3.6, 3.0, and 2.0, while the number of markers on an average chromosome of 150 cM increases from 20 to 200 (table 4). At a relatively low density of 20 markers, the average number of false-positive errors is ~.0017, just below the intended type I error, when LOD score 3.6 is used. For LOD score 3.0, the average number of false-positive markers, .0058, is comparable to .0057, the type I error obtained from the earlier simulation under the null hypothesis. Now, as the number of markers increases from 20 to 200, the average number of false-positive errors increases accordingly, from .0017 to .0176, for LOD score 3.6, and from .0058 to .0618, for LOD score 3.0; if LOD score 2.0 is used, the average number of false-positive errors reaches .5181, with 200 markers. Given such a strong association with the number of markers, it is natural to conclude that, as an index for measurement of genomewide significance, the expected number of false-positive errors is less preferable than the type I error.

Table 4

Average Number of False-Positive Errors under the Null Hypothesis as ε Increases from 20 to 200, on an Average Chromosome of 150 cM

No. of Markers	Average No. of False Positive Errors, for		
	Lod Score 3.6	Lod Score 3.0	Lod Score 2.0
20	.0017	.0058	.0551
40	.0034	.0134	.1056
60	.0051	.0181	.1571
80	.0061	.0238	.2067
100	.0081	.0307	.2665
120	.0107	.0355	.3079
140	.0123	.0429	.3635
160	.0133	.0449	.4017
180	.0170	.0558	.4647
200	.0176	.0618	.5181

Discussion

The principal conclusions to be derived from the Monte Carlo simulation experiments discussed above are fivefold. (1) It appears that the reliability of the LOD-score-3.0 criterion and that of the LOD-score-3.6 criterion are rather comparable in testing and prediction, in the presence of one or two major genes; however, with regard to ensuring that the type I error at the chromosomal level is .0021, the use of LOD score 3.6 appears to be more appropriate than the use of LOD score 3.0, even though the type I error is influenced by the sample size and by the chromosome sizes. Hence, this simulation supports the use of LOD score 3.6 as a conservative way of ensuring that the genomewide significance under the null hypothesis is .05. (2) In disease-gene discovery, the presence of two or more major genes substantially reduces the reliability—that is, there is reduced TPP and increased FPP, even though the reduction in power is modest. To improve the utility of linkage-based mapping, it seems necessary, for detection of multiple genes on the same chromosome, to search for methods that are more powerful than the use of just a simple cutoff. (3) Limited simulations have also shown that reliability is rather sensitive to the choice of ϵ, which determines the precision of the prediction. In balancing the precision of the predicted genomic region versus adequate reliability in the prediction of such a region, one may consider choosing an ϵ that is ~10 cM if a single gene is postulated and choosing an ϵ that is ~20 cM if two genes are postulated. (4) Although an increase in the number of markers—for example, from 20 to $\geqslant$40 markers/chromosome—improves power and reliability, having an excessively large number of markers—for example, >40 markers/chromosome—results in only a small additional return. (5) Compared with an increase in the number of markers, an increase in sample size—that is, having either more family members or more families—is probably a more desirable way to increase power, if a single-gene alternative is used. If two or more disease genes are suspected on the same chromosome, then the power improvement that is due to an

increase in sample size is noticeable, but the improvement in either TPP or FPP is rather modest. This result suggests that it is important to increase sample sizes of homogeneous groups, if the current testing procedure is to used; alternatively, however, one may develop a new approach that explicitly incorporates the presence of multiple disease genes. It is anticipated that the linkage analysis with an appropriate tool would have improved power with an increase in sample size of even heterogeneous groups.

Although these conclusions appear to be sensible, it is important to recognize the context in which these conclusions are drawn. First, each experiment simulates a sample of fully informative meioses, in which all recombination fractions are directly estimable by averaging the recombinant indicators. Although such meiotic data are obtainable in controlled experiments, they are rarely available in human studies. With observational human data, the situation considered by Sawcer et al. (1997), meioses generally have to be inferred from pedigree data indirectly, under a set of additional assumptions. In practice, some of these assumptions may be violated, leading to "noisy" estimators of recombination fractions and thereby reducing the signal/noise ratio. Consequently, it is possible that associations identified in the ideal experiment may differ somewhat under more-realistic sampling circumstances.

Second, most simulation experiments assume a particular placement of the putative disease gene in the map and also a mixture of linked and unlinked meioses. This mechanism of recombination process may be different from the actual process, which could depend on pedigree structures, penetrance, disease-allele frequency, and/or marker-allele frequency.

Third, the procedure described above predicts that one or more disease genes may be contained in the interval, from $-\epsilon$ to $+\epsilon$, around the marker with the maximum LOD score, and that the choice of the constant interval is somewhat arbitrary. To relax this restriction, one may use the LOD-score values or estimated recombination fractions around the marker of the maximum LOD score and then, making use of empirical data, predict those intervals; alternatively, one may also consider modeling patterns of recombination fractions rigorously, as a way of determining locations of disease genes, an approach that will be explored in the future.

Fourth, this simulation setup assumes knowledge about the presence of one or two (or more) disease genes on the same chromosome. In practice, such knowledge is not available, and the absence of such knowledge presumably would reduce the power of simple test statistics. Overcoming this problem calls for new methodologies that retain power in the presence of multiple disease genes. Although developing methods to address multiple disease genes is necessary and important, it is also important to point out that the presence of multiple genes on different chromosomes may be more common than multiple disease genes on a single chromosome. Of course, the genetic heterogeneity across chromosomes has been implicitly addressed by allowing unlinked meioses in simulations and has been shown to be less problematic.

Fifth, the simulation setup generally assumes that there is no interference, since the latter is believed to play a limited role in the human genome; however, interference has been shown in the yeast genome, and its presence in the human genome could affect the dependence structure of recombination events and, hence, alter aspects of our evaluation. For this reason, we followed the simulation procedure described above and simulated correlated recombination events with a dependence parameter, which took values of .00–.30, corresponding to the range from no interference to strong interference. Table 5 lists estimated chromosome-specific type I errors when either LOD score 3.6 and LOD score 3.0 is used, when the interference parameter has values of .00, .05, .10, .20 and .30. Clearly, the presence of interference inflates the type I error, and, depending on the magnitude of the interference, this inflation could be substantial.

Bearing in mind the specific setup of these Monte Carlo simulation experiments, we have learned some lessons beyond the conclusions discussed above. The first lesson concerns the nature of pointwise *P* values. Typically, a pointwise *P* value corresponding to a LOD score—for example, LOD score 3.0—is on the order 10^{-4}, which is an impressive number to report. In our limited Monte Carlo experiments, we observed LOD scores >3.0 that were present somewhere along the genome, with a frequency much higher than 10^{-4}. In the statistical literature, this problem is known as a multiple-comparison problem (see Hsu 1996). Without an appropriate adjustment for multiple comparison, the pointwise significance levels are misleading and should not be reported.

As a second lesson, this simulation study allows us to

Table 5

Type I Errors for LOD scores 3.6 and 3.0, in the Presence of Interference, from 0 (Absence of Interference) to .05 (Weak Interference) to .10 (Modestly Weak Interference) to .20 (Modest Interference) to .30 (Strong Interference)

Interference Parameter	Type I Error, for	
	LOD Score 3.6	LOD Score 3.0
.00	.00196	.00597
.05	.00244	.00681
.10	.00344	.00966
.20	.01459	.03268
.30	.05468	.10466

appreciate the gain from point estimates of recombination fractions, besides LOD scores. It is recognized that LOD scores are generally used to search for linkage signals. A large LOD score at a marker locus implies either that the recombination fraction is deviating from .50 or that the number of informative meioses at that locus is substantial. Hence, the recombination fraction could be informative about the location of the disease gene, in a situation in which the marker is fairly close to the putative disease gene and yet is less informative than neighboring markers. For example, consider a single-nucleotide–polymorphism marker that is near the disease gene. This marker may be less heterozygous than its neighboring markers. In such cases, estimated recombination fractions could be used in conjunction with LOD scores.

The third lesson is related to the fine scale–mapping strategy. A typical mapping study aims to identify marker loci with significant signals and with the highest LOD score and to map genes at a fine scale (e.g., ~1 cM) around those loci, using, for example, linkage-disequilibrium analysis. Our simulation has shown, however, that use of narrow intervals, which may be necessary for linkage-disequilibrium mapping, may have a good chance of missing the true disease-gene location. On the other hand, the power for fine-scale mapping by linkage-disequilibrium analysis may be limited for large intervals. Balancing these opposing considerations, one may have to consider either an increase in sample size, to narrow the intervals, or an increase in the number of markers, to improve the linkage-disequilibrium signals, along with alternative strategies—such as the use of candidate genes in targeted regions—for efficient fine-scale mapping.

The fourth lesson concerns the relationship between LOD scores and Wald-test statistics. It is well known that the LOD score (or, equivalently, the likelihood ratio) is asymptotically equivalent to the Wald-test statistic for likelihood-based inference (e.g., see Cox and Hinkley 1986). It is also known that these two test statistics could differ in some situations (Moolgavkar and Vezon 1987). In the current context of linkage analysis, we compared the Wald-test statistic versus the LOD score, on the same scale (fig. 7). Interestingly, the pattern of LOD scores and that of Wald-test statistics are generally comparable, except at the boundary value of 0 recombination fraction. Other examples may be simulated on our Website.

Finally, the results of the study of genomewide significance in the present report also have implications for genomewide searches for linkage-disequilibrium mapping, for studies of loss of heterozygosity, and for studies with a large gene-expression array. Rather similar to linkage analysis, linkage-disequilibrium analysis and loss-of-heterozygosity analysis also take advantage of

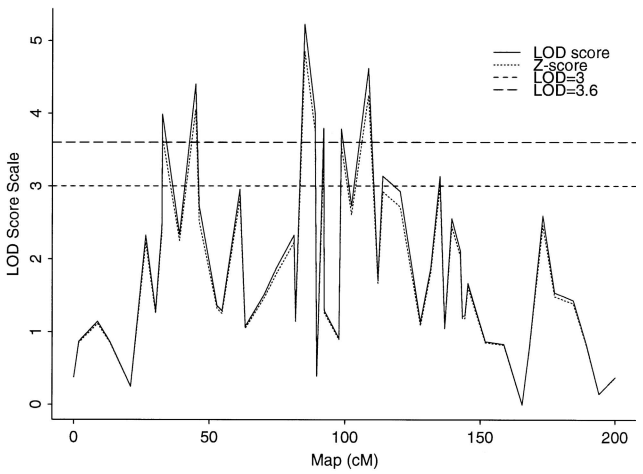

Figure 7 Typical pattern of LOD scores and Z-scores from a simulated genome scan with the mixing proportion $\lambda = .5$ (other examples may be simulated online at the Fred Hutchinson Cancer Research Center QGE Website).

the linear genomic structure and compute a vector of correlated test statistics for a group of linearly ordered markers along the genome, in addition to a vector of estimated parameters that quantify the associations of interest. The unique feature, however, is that the dependencies induced by the linkage-disequilibrium analysis and by the loss-of-heterozygosity analysis are different from those induced by linkage analysis, thereby having an impact on the assessment of the related genomewide significance; for example, the correlation of linkage-disequilibrium signals is weaker than that of linkage signals between adjacent marker loci, which may result in a further increase in LOD-score cutoff values, if the maximum LOD score is used as the test statistic.

On the other hand, the analysis of a large array of expressed genes presents a different challenge to the assessment of the "genomewide" significance. Indeed, functional genes—for example, those on the current Affymetrix array of 6,800 genes—are chosen primarily for their functional significance, particularly with regard to their regulation during the cell-cycle process. Even though all these genes are located somewhere in the genome, the linear genomic structure is of less importance, since the primary interest is in their association with a certain phenotype. An important property, however, is that the functionality of these genes is organized via an underlying network with many pathways by which genes may penetrate to phenotypes of interest (L. Hood, in the ceremony for the opening of the Thomas Building at the Fred Hutchinson Cancer Research Center). Those genes, on the same pathway, tend to be highly correlated, whereas genes not sharing the pathway may be much

less correlated or even uncorrelated. Unfortunately, the underlying network structure is unknown, prohibiting any straightforward grouping of thousands of functional genes.

Hence, in much of the general context of genome-scan studies, it appears that the use only of maximum LOD scores (or an equivalent statistic) is probably too simplistic. To be fully efficient in the testing of the null hypothesis, one probably has to use the test statistic that has power under a broad range of alternatives. Meanwhile, to ensure the greatest reliability, one may need to have a flexible algorithm to search the genome for all possible leads, in the hope of improving TPP and reducing FPP, under a broad range of alternatives. Motivated by this consideration, we are currently developing a two-step procedure: in the first step, one tests the null hypothesis; in the second step, one searches for positive leads for disease genes. Separate statistics are chosen for each steps, and each is optimized to achieve the desired efficiency. The results of this work will be provided in a future report.

Acknowledgments

The authors would like to thank Dr. Robert Elston for some helpful discussion. This work is supported, in part, by National Institutes of Health grants CA53996, CA72964, and AG14358.

Electronic-Database Information

Fred Hutchinson Cancer Research Center QGE, http://lynx.fhcrc.org/~qge

References

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al (1996) Accessing genetic information with high density DNA arrays. Science 274:610–614

Cox DR, Hinkley DV (1986) Theoretical statistics. Chapman & Hall, New York

Curtis D (1996) Re: genetic dissection of complex traits. Nat Genet 12:356–357

Feingold E, Brown PO, Siegmund D (1993) Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–251

Hsu JC (1996) Multiple comparisons: theory and methods. Chapman & Hall, London

Lander ES, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Moolgavkar SH, Vezon DJ (1987) Confidence regions in curved exponential families: application to matched case-control and survival studies with general relative risk function. Ann Stat 15:346–359

Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277-318

——— (1998) Significance levels in complex inheritance. Am J Hum Genet 62:690–697

Rao DC (1998) CAT scans, PET scans, and genomic scans. Genet Epidemiol 15:1-18

Rothman KJ, Greenland S (1998) Modern epidemiology. Lippincott-Raven, Philadelphia

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D (1997) Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077–1082

Witte JS, Elston RC, Schork N (1996) Re: genetic dissection of complex traits. Nat Genet 12:355–356