Am. J. Hum. Genet. 68:1219–1228, 2001

A Confidence-Set Approach for Finding Tightly Linked Genomic Regions

Shili Lin, James A. Rogers, and Jason C. Hsu

Department of Statistics, The Ohio State University, Columbus

As more studies adopt the approach of whole-genome screening, geneticists are faced with the challenge of having to interpret results from traditional approaches that were not designed for genome-scan data. Frequently, two-point analysis by the LOD method is performed to search for signals of linkage throughout the genome, for each of hundreds or even thousands of markers. This practice has raised the question of how to adjust the significance level for the fact that multiple tests are being performed. Various recommendations have been made, but no consensus has emerged. In this article, we propose a new method, the confidence-set approach, that circumvents the need to correct for the level of significance according to the number of markers tested. In the search for the gene location of a monogenic disorder, multiplicity adjustment is not needed in order to maintain the desired level of confidence. For complex diseases involving multiple genes, one needs only to adjust the level of significance according to the number of disease genes—a much smaller number than the number of markers in a genome screen—to ensure a predetermined genomewide confidence level. Furthermore, our formulation of the tests enables us to localize disease genes to small genomic regions, an extremely desirable feature that the traditional LOD method lacks. Our simulation study shows that, for sib-pair data, even when the coverage probability of the confidence set is chosen to be as high as 99%, our approach is able to implicate only the markers that are closely linked to the disease genes.

Introduction

With the whole-genome–scan approach becoming almost a routine matter, geneticists are faced with the challenge of having to interpret results from traditional approaches, as well as having to find statistical methods that are more appropriate in this setting. As a first path to identifying linked genomic regions, two-point analysis by the LOD method (Morton 1955) is usually performed for each marker (Ott 1999, p. 114), with the number of markers for a genome-scan study ranging from a few hundred to a couple thousand. Each analysis (i.e., the analysis for each marker m) amounts to testing the null hypothesis of no linkage, $H_{0m}:\theta_m = 1/2$, versus the alternative hypothesis of linkage, $H_{am}:\theta_m < 1/2$, where θ_m is the recombination fraction between a disease gene and marker m.

The original recommendation, which declares significance for linkage when the LOD score is $\geqslant 3$ (Morton 1955) and which was not proposed for the purpose of the genome scan, has been the focal point of an ongoing debate about the mapping of complex traits. During the 1980s, rapid progress in molecular genetics provided a large number of RFLP markers for gene mapping. In light of the potential for inflated type I error (i.e., the declaration that linkage exists when, in fact, it is absent) when multiple markers are tested by the LOD method, several different cutoffs were proposed, based on diverse arguments, including exact calculation, Bonferroni correction, and Bayesian formulation (Kidd and Ott 1984; Thompson 1984; Ott 1985; Edwards and Watt 1989; Risch 1991). With many studies now based on as many as thousands of microsatellite markers throughout the genome, the problem caused by multiple testing becomes more severe, prompting the current ongoing debate (Lander and Kruglyak 1995; Curtis 1996; Witte et al. 1996; Sawcer et al. 1997; Morton 1998; Zhao et al. 1999).

The central issue of the debate is how to reduce the type I error rate to an acceptable level genomewide yet still be able to detect signals for linkage when multiple tests are performed. Many of the proposed ideas involve controlling the *traditional* type I error associated with the LOD method. We refer to the type I error associated with the LOD method (i.e., the inference of false linkage) as the "traditional" type I error, to distinguish it from the type I error that is associated with the procedure that we are proposing in this article. The simple LOD-score cutoff of 3 is generally regarded as not being stringent enough to avoid too many false declarations of linkage, for complex traits, when multiple tests are performed. There were several cutoff values, for different scenarios, given by Lander and Kruglyak (1995), the most often quoted being 3.6 (e.g., see Zhao et al. 1999). Others have proposed cutoff values between 3

Received January 22, 2001; accepted for publication March 2, 2001; electronically published April 13, 2001.

Address for correspondence and reprints: Dr. Shili Lin, Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247. E-mail: shili@stat.ohio-state.edu

0002-9297/2001/6805-0016$02.00

and 3.6, all in the hope of maintaining a genomewide error rate of $\leq 5\%$ (e.g., see Sawcer et al. 1997; Morton 1998). Apart from the traditional type I error, other measures of genomewide significance include the false-discovery rate (Drigalenko and Elston 1997), false-positive–prediction error (FPP) and the related reliability index (Morton 1998), and expected number of false-positive errors (Zhao et al. 1999). Similar measures are discussed with respect to our formulation in the present study.

In the early days of human linkage mapping, before the development of DNA variants, the number of genetic markers (mostly blood groups and enzymes) was limited; only as many as 60 classic marker loci were available (Ott 1999). Because of the limited availability of markers when the LOD method was proposed, the strategy was to find any marker that was linked to the disease locus. Such a marker, if it exists, may be only loosely linked to the disease locus (with a recombination fraction [θ] between the two loci of, say, .3). A distance of $\theta \leq .3$ is generally regarded as mappable (Risch 1991; Drigalenko and Elston 1997); hence, any marker located less than that distance from the disease locus has a good chance of being identified as linked, provided that there is sufficient statistical power. In fact, for data from fully informative meioses, for example, any marker residing on the same chromosome as that harboring the disease locus will test positive for linkage asymptotically, owing to the fact that the LOD score will approach infinity as the amount of data increases.

In the context of genome-scan studies with markers spanning the genome at a density of, say, 10 cM, many markers will provide significant signals for linkage even when the type I error is controlled to be less than 5% (or when any of the aforementioned measures are being controlled). We suggest that control of traditional type I error should not be the focus in genome-scan studies. Identification of all markers that are mappable should no longer, in our view, be the primary goal, now that dense maps of markers are available. We believe that the focus should shift to the localization of disease genes to small chromosomal regions, even at the stage of preliminary genome screening.

The present study proposes a new statistical procedure for identification of tight linkage (with, say, ≤ 5 cM between the marker and disease loci). The approach is designed to identify only the markers that are within a specified (small) distance d_0 (in cM) of a disease locus. The idea is to construct a confidence set of markers, with coverage probability at a predetermined level, p; in other words, we want to compile a set of markers that will allow us to say, with confidence $100p\%$, that every disease gene is within d_0 of one of the markers in the set. The distance d_0 is usually chosen to be such that, asymptotically, for a map of equally spaced markers, only one unique marker for each disease gene is included in the confidence set, provided that the disease locus is not exactly halfway between two markers. This ensures that, with sufficient data, only markers that are tightly linked to a disease locus are included in the confidence set, and our method thereby effectively identifies small genomic segments that may contain disease genes. Furthermore, multiplicity adjustment for the number of markers is not needed, even when thousands of markers are tested one by one.

We would like to mention, in passing, that the problem that we are addressing in the present study is different from the problem that was discussed by Elston and Lange (1975) and Lange and Boehnke (1982), although the two problems are related. The issue in those studies was the determination of the number of random markers that were needed to cover the genome such that the probability was high that there was a marker within a certain distance of a disease locus. One may view our problem as the exact opposite. Here, we have a fixed map of markers covering the genome, and we are interested in finding disease genes that are located, with a high level of confidence, within a certain distance of the nearest markers.

Methods

Construction of a Confidence Set

Suppose that we have M markers along the chromosomes, with the largest intermarker distance denoted by d. Let $d_0 = d/2$. Then, for a disease locus in the genome, there exists a marker that is within d_0 of the disease gene. The goal is to find a set of markers, A, that will, with probability p, include at least one such marker. Such a set of markers is referred to as a "confidence set" with coverage probability p. If all the markers are equally spaced and the disease locus is not located exactly halfway between two markers, then there is a unique marker, m^*, that is within d_0 of the disease gene. Although our procedure works under the general scenario, we focus our discussion below on the somewhat restricted setting of a unique marker, for the purpose of presentation. Thus, we want to find A such that $P(m^* \in A) \geq p$.

By the duality of confidence set and hypothesis testing (Bickel and Doksum 1977), constructing a confidence set is equivalent to testing the following hypotheses for each marker m:

$$H_{0m}: d_m \leq d_0 \text{ vs. } H_{am}: d_m > d_0 ,$$

where d_m is the true but unknown distance between a disease locus and marker m. Since we are dealing with a two-locus analysis, it is more natural to represent distance between two loci in terms of θ rather than in terms of

genetic distance. Let θ_m and θ_0 be the θ values corresponding to d_m and d_0, respectively. The equivalent hypotheses are: H_{0m}: $\theta_m \leqslant \theta_0$ and H_{am}: $\theta_m > \theta_0$. The generalized likelihood-ratio test statistic is

$$\lambda_m = \frac{\sup_{\theta_m \leqslant \theta_0} L(\theta_m)}{\sup_{\theta_m \in [0,1/2]} L(\theta_m)} = \begin{cases} 1 & \text{if } \hat{\theta}_m \leqslant \theta_0 \\ \dfrac{L(\theta_0)}{L(\hat{\theta}_m)} & \text{if } \hat{\theta}_m > \theta_0 \end{cases},$$

where $\hat{\theta}_m$ is the maximum-likelihood estimate (MLE) of an (assumed) unimodal likelihood function $L(\theta_m)$. We need to find c_α ($\leqslant 1$) such that

$$\sup_{\theta_m \leqslant \theta_0} P_{\theta_m}\left(\frac{L(\theta_0)}{L(\hat{\theta}_m)} < c_\alpha\right) = \alpha ,$$

where $\alpha = 1 - p$ is the type I error of our test. Then $A = \{m:\lambda_m \geqslant c_\alpha\}$ is a confidence set with coverage probability $P(m^* \in A) = P_{\theta_{m^*}}(\lambda_{m^*} \geqslant c_\alpha) \geqslant 1 - \alpha = p$.

In the following two subsections, we derive confidence sets for two family types—the phase-known (PK) double backcross and the phase-unknown (PU) double backcross—to illustrate our procedure. In both family types, each offspring has one doubly homozygous parent (not informative for linkage) and one doubly heterozygous parent. The phase of the doubly heterozygous parent is known in the PK family type but is unknown in the PU family type. These two data types are frequently used by researchers when they are investigating exact properties of proposed procedures (e.g., see the report by Ott [1999]). We assume, in the derivation of the formulae below, that n families with two offspring (sib pairs [SPs]) per family are available.

PK Sib-Pair Data

For marker m, let θ_m be the true recombination fraction between the marker and the disease locus. The PK double backcross allows unambiguous classification of recombination events. Let $S_m = \sum_{i=1}^n X_i$ be the number of recombinants in the total of $2n$ meioses, where X_i is the number of recombinants in SP i. Let $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ be a realization of $\boldsymbol{X} = \{X_1, \cdots, X_n\}$. Then the likelihood for θ_m is

$$L(\theta_m \mid \boldsymbol{x} = \{x_1, \cdots, x_n\}) = \prod_{i=1}^n \binom{2}{x_i} \theta_m^{x_i}(1-\theta_m)^{2-x_i} \propto \theta_m^{s_m}(1-\theta_m)^{2n-s_m} ,$$

where s_m is a realization of S_m. Thus S_m is a sufficient statistic for θ_m, and the generalized likelihood-ratio test amounts to rejection of the null hypothesis (tight linkage) when S_m is large. To control the type I error of our test to be α, we need to find c_α such that

$$\sup_{\theta_m \leqslant \theta_0} P_{\theta_m}(S_m > c_\alpha) = \alpha .$$

We wish to reiterate that controlling for type I error in our test differs from controlling the traditional type I error of false linkage in the LOD method, since the null hypotheses being tested in the two situations are different; in fact, what we are controlling in our test is the coverage probability of a confidence set.

Since larger θ values lead to greater chances of recombination,

$$\begin{aligned} \sup_{\theta_m \leqslant \theta_0} P_{\theta_m}(S_m > c_\alpha) &= P_{\theta_0}(S_m > c_\alpha) \\ &= \sum_{s_m > c_\alpha} \binom{2n}{s_m} \theta_0^{s_m}(1-\theta_0)^{2n-s_m} \\ &\approx 1 - \Phi\left[\frac{c_\alpha - 2n\theta_0}{\sqrt{2n\theta_0(1-\theta_0)}}\right] , \end{aligned}$$

where Φ is the cumulative distribution function of the standard normal variate. Although one may find the exact type I error by summing over the binomial probabilities, one can also approximate this probability by using the central-limit theorem (CLT), as shown in the last expression of the formulae above, as long as $2n\theta_0$ is sufficiently large (say, $\geqslant 10$). Using the normal approximation formula, we find that

$$c_\alpha = 2n\theta_0 + \Phi^{-1}(1-\alpha)\sqrt{2n\theta_0(1-\theta_0)} .$$

The confidence set, $A = \{m:s_m \leqslant c_\alpha\}$, has coverage probability of at least $1 - \alpha$:

$$\begin{aligned} P(m^* \in A) &= P_{\theta_{m^*}}(S_{m^*} \leqslant c_\alpha) \\ &\geqslant 1 - P_{\theta_0}(S_{m^*} > c_\alpha) \\ &= 1 - \alpha ; \end{aligned}$$

that is, we are at least $100(1-\alpha)\%$ confident that the set A includes m^*, the only marker that is less than d_0 from the disease.

PU Sib-Pair Data

For marker m, let θ_m be as defined above in the PK data type. Let X_i be the number of recombinants under one of the two phases of the heterozygous parent, for sib pair i. Let $S_m = \sum_{i=1}^n \min\{X_i, 2 - X_i\}$ be the number of families that has exactly one recombinant under one of the phases. Under the assumption of linkage equilibrium, the two phases are equally likely; so the likelihood function for θ_m is

$$L(\theta_m|\boldsymbol{x} = \{x_1, \cdots, x_n\}) = \prod_{i=1}^{n} \frac{1}{2}\binom{2}{x_i}[\theta_m^{x_i}(1-\theta_m)^{2-x_i} + \theta_m^{2-x_i}(1-\theta_m)^{x_i}] \propto [\theta_m(1-\theta_m)]^{s_m}[\theta_m^2 + (1-\theta_m)^2]^{n-s_m} ,$$

where s_m is a realization of S_m. Thus, S_m is a sufficient statistic for θ_m. As an aside for the more theoretically inclined reader, S_m is, in fact, minimal sufficient. Furthermore, similar sufficient statistics can be found for the PU double backcross with k offspring, for $k > 2$ (Rogers et al. 2001).

The generalized likelihood-ratio test again amounts to rejection of the null hypothesis when S_m is large. As shown in Appendix A,

$$\sup_{\theta_m \leqslant \theta_0} P_{\theta_m}(S_m > c_\alpha) = P_{\theta_0}(S_m > c_\alpha) .$$

Since the distribution of S_m can be derived exactly, one may perform exact calculation of $P_{\theta_0}(S > c_\alpha)$ to find a c_α such that the type I error is, at most, α. Alternatively, since S_m is the sum of independent and identically distributed random variables, we can approximate the probability by the CLT, leading to a satisfactory estimate of the critical value,

$$c_\alpha = 2n\theta_0(1-\theta_0) + \Phi^{-1}(1-\alpha)\sqrt{2n\theta_0(1-\theta_0)[1-2\theta_0(1-\theta_0)]} ,$$

when $2n\theta_0(1-\theta_0)$ is $\geqslant 10$. The resulting confidence set $A = \{m : S_m \leqslant c_\alpha\}$ can be shown, as in the previous subsection, to have a coverage probability of at least $1-\alpha$.

It is interesting to note that a statistic similar to S_m was proposed, by Bernstein (1931), for estimation of θ. This statistic, $Y = \sum_{i=1}^{n} X_i(2 - X_i)$, was, however, shown, by Fisher (1935), to be less efficient than the MLE, owing to the waste of linkage information. The statistic S_m, on the other hand, is fully efficient.

Expected Number of Markers in Confidence Set

The expected number of false-positive errors is conventionally defined as the expected number of markers that are unlinked to the disease but that are inferred, by a statistical test, to be linked. This number may be used as a measure of genomewide significance (Zhao et al. 1999). In our formulation for finding tightly linked genomic regions, a similar measure is the expected number of markers included in the confidence set. The expected number of false inclusions, defined as the number of markers that are more than d_0 from any disease locus but that are included in the confidence set, is another closely related measure. If there is a unique marker within d_0 of each disease locus, then, ideally, we would like the confidence set to be composed of these markers only—that is, without any false inclusion. However, for a finite data set, by controlling the coverage probability to be p, we would expect the number of markers contained in the confidence set to be greater than the number of disease genes. Therefore, a confidence set with a smaller number of false inclusions is considered to be better than one with a greater number of false inclusions, provided that both sets have the same coverage probability. For a confidence set constructed through use of the procedure proposed in this article, the expected number of false inclusions is a function of the sample size, as is shown in the next subsection.

Sample Size and False Inclusions

Recall that, for the PK family data, when the critical value is chosen to be

$$c_\alpha = 2n\theta_0 + \Phi^{-1}(1-\alpha)\sqrt{2n\theta_0(1-\theta_0)} ,$$

the formula for controlling the coverage probability (type I error of our test),

$$\sup_{\theta \leqslant \theta_0} P_\theta(S_m > c_\alpha) = \alpha ,$$

holds, approximately; that is, the coverage probability of $A = \{m : S_m \leqslant c_\alpha\}$ is at least $p = 1-\alpha$. We want to find the smallest sample size n (number of families) such that, with probability q, a marker located at a distance θ ($>\theta_0$) from the disease locus will not be included in the confidence set. This implies that $q = P_\theta(S_m > c_\alpha)$. Application of normal approximation leads to the following sample-size formula:

$$n = \left\lceil \left[\Phi^{-1}(1-\alpha)\sqrt{2\theta_0(1-\theta_0)} - \Phi^{-1}(1-q)\sqrt{2\theta(1-\theta)}\right]^2 \Big/ 4(\theta-\theta_0)^2 \right\rceil , \qquad (1)$$

where $\lceil\ \rceil$ is the ceiling of the number being bracketed (i.e., the smallest integer that is greater than or equal to the number). The probability q is the power of our test. Requiring a smaller number of expected false inclusions implies a greater power, which, in turn, requires a larger sample size, as confirmed by the sample-size formula above.

Similarly, we obtain an approximation formula for

Table 1

Number of PK Families Necessary to Satisfy the Specifications.

d[a] (cM)	No. of Families at Power =[b] 99%	95%	90%	85%	80%
6	6,643	4,772	3,901	3,363	2,963
7	1,820	1,292	1,048	897	786
8	878	618	497	424	369
9	533	371	297	252	219
10	365	253	201	170	147
11	271	186	148	124	107
12	211	144	114	96	82
13	171	117	92	77	66
14	143	97	76	63	54
15	122	82	64	53	45[c]

NOTE.—Numbers are computed from equation (1), based on normal approximation.

[a] $d_0 = 5$ cM.

[b] For all levels of power, the coverage probability of the confidence set is controlled to be 99%.

[c] The sample size for this particular setting may not be estimated very accurately, because of normal approximation with a small expected cell count.

calculation of the number of PU families that are needed in order to meet the desired specifications:

$$n = \left\lceil \left\{ \Phi^{-1}(1-\alpha)\sqrt{2\theta_0(1-\theta_0)[1-2\theta_0(1-\theta_0)]} - \Phi^{-1}(1-q)\sqrt{2\theta(1-\theta)[1-2\theta(1-\theta)]} \right\}^2 \Big/ 4[\theta(1-\theta)-\theta_0(1-\theta_0)]^2 \right\rceil . \quad (2)$$

Setup of a Simulation Study

To compare and contrast results from the LOD method and the confidence-set approach, as well as to investigate the influences that several mapping parameters have on the outcomes of the procedure proposed in this article, we performed a simulation study with data from PK and PU SPs. In most of the simulations, 250 SPs were used. Each chromosome was usually assumed to be covered by 30 equally spaced markers 10 cM apart. The disease gene was assumed to be located between markers 15 and 16, 1 cM from the former. The coverage probability was controlled to be 99%. For each parameter under investigation, values that differ from these general settings were specified when they occurred. A total of 1,000 replications were used for each simulation.

Results

Sample-Size Determination

Table 1 shows the minimum sample sizes required for various levels of power, computed from equation (1), for the PK-family data type. We assume that $d_0 = 5$ cM and we control the confidence set to have 99% coverage probability. Each entry in the table gives the sample size needed for a marker, located at distance d ($>d_0$) from the disease locus, not to be included in the confidence set with the specified probability (power). For example, if one desires that, with probability 99%, any marker located more than d_0 from the disease locus should not be included, then a few thousand PK SPs are needed. On the other hand, if an equally spaced map with a 10-cM marker density is being used, then any marker that is not one of the two markers flanking the disease locus is located $\geqslant$10 cM from the disease locus. This implies that, with high probability, only a couple hundred SPs are needed for the confidence set to exclude markers that do not flank the disease gene. Table 2 gives the results, using equation (2), for the PU-family data type. As expected, more PU families than PK families are required in order to achieve the same specifications of powers and coverage probabilities. The increases in the numbers of families are 12%–29%, with an average increase of ~19%.

The remainder of this section describes results from the simulation study. In many of the simulations, 250 PK SPs and a 10-cM density map were used. As discussed above and shown in table 1, such a data set would provide sufficient power to limit false inclusions in confi-

Table 2

Number of PU Families Necessary to Satisfy the Specifications.

d[a] (cM)	No. of Families at Power =[b] 99%	95%	90%	85%	80%
6	7,457	5,365	4,390	3,788	3,341
7	2,064	1,470	1,194	1,025	899
8	1,006	710	574	490	428
9	616	432	347	295	257
10	426	297	238	202	175
11	319	221	176	149	129
12	251	173	138	116	100
13	205	141	112	94	81
14	172	118	94	79	68
15	148	101	80	67	58

NOTE.—Numbers are computed from equation (2), based on normal approximation.

[a] $d_0 = 5$ cM.

[b] For all levels of power, the coverage probability of the confidence set is controlled to be 99%.

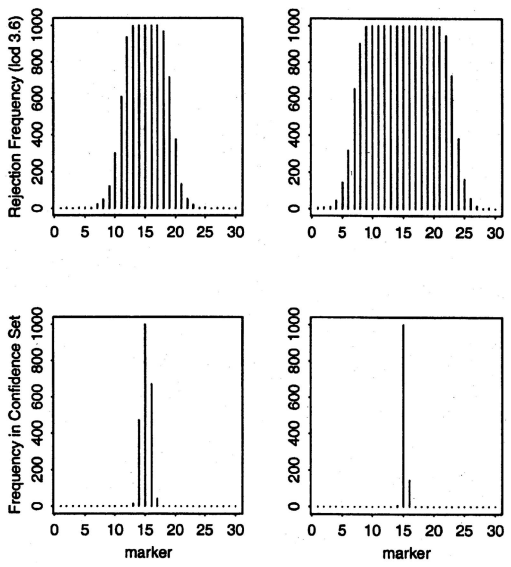

Figure 1 Comparison of results from the LOD method and the confidence-set approach. The top panel shows the results of using the LOD method with a cutoff of 3.6. The height of each vertical line represents the inferred frequency (per 1,000 replications) of linkage of the marker (i.e., the frequency with which the null hypothesis of no linkage was rejected). The bottom panel shows the results from the confidence-set approach with 99% coverage probability. The height of each vertical line represents the frequency with which the marker was included in the confidence set. The two plots on the left are for data with 50 PK SPs, whereas the two plots on the right are for data with 250 PK SPs.

dence sets and therefore was chosen to be the primary data structure in our simulation.

Confidence Sets, and Inference of Linkage by the LOD Method

We compared conclusions drawn from analyses based on the LOD method versus conclusions from analyses based on the confidence-set approach. Note that these two methods are based on the testing of "reversed" hypotheses; hence, they are not directly comparable in terms of type I error or power. What we are comparing are the implications, in terms of finding regions containing disease genes, drawn from these two procedures, especially their abilities to narrow the linkage regions when more data are used. Results for two data sets are shown in figure 1. The LOD method, with a cutoff of 3.6, rejected the null hypothesis of no linkage, for many markers, even those not very close to the disease (fig. 1, *top two plots*). In fact, as data accumulated (from 50 SPs to 250 SPs), most of the markers located on the same chromosome as the disease were inferred to be linked. Results from the confidence-set approach picked up far fewer markers (fig. 1, *bottom two plots*) since this approach is designed to detect tight linkages only. For the data sets with 50 SPs, the only marker (marker 15) <5 cM from the disease locus was inferred to be in the confidence set in all the replications. In ~50% of the replicates, one more marker (either marker 14 or 16) was also included in the confidence set. In only ~1% of the replicates was a marker as far as almost 20 cM from the disease included in the confidence set. When the number of SPs was increased to 250, the confidence set contained only marker 15 in 85% of all the replications. Marker 16 was 9 cM from the disease locus and was included in 144 (~15%) of the confidence sets. Marker 14, on the other hand, was 11 cM from the disease locus and was included in six (~1%) of the confidence sets. These empirical results correspond well with the theoretical results given in table 1.

Figure 2 shows the distributions of the number of markers inferred to be linked (LOD method) or tightly linked (confidence-set approach). These histograms summarize the results for all markers jointly rather than individually. With just 50 SPs, typically 9 or 10 markers (spanning a region of ~100 cM) were inferred to be linked to the disease (fig. 2, *top left*). When more data became available, the null hypothesis (which was not true) was rejected much more often, implicating an ~170–180-cM region (fig. 2, *top right*). These results demonstrate that, with the LOD method, although one

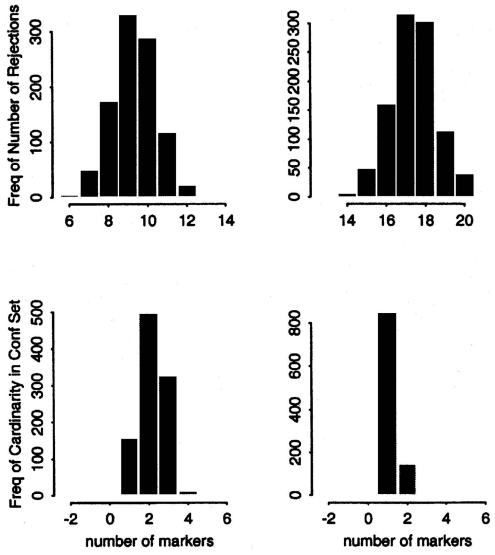

Figure 2 Distributions of numbers of markers inferred to be linked under the LOD method and the confidence-set approach. The top panel shows the results of using the LOD method with a cutoff of 3.6. The histograms display the distributions (per 1,000 replications) of the number of markers inferred to be linked. The bottom panel shows the results from the confidence-set approach with 99% coverage probability. The histograms display the distributions of the number of markers included in the confidence set. The two data sets are as described in figure 1.

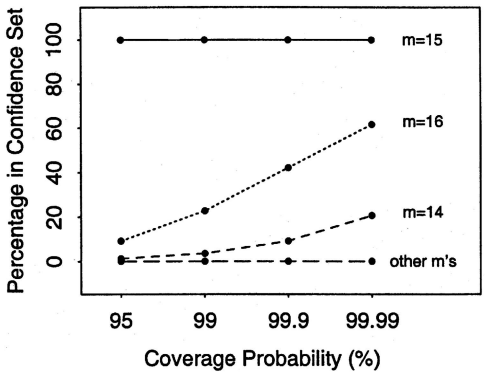

Figure 3 Effects of coverage probability on the number of markers in the confidence set. Each set of dots connected by lines represents the relative frequency at which the marker was included in the confidence set, over a range of coverage probabilities. The "Coverage Probability" axis is not drawn to the scale and should simply be interpreted as representing four categories.

is able to detect significant signals for linkage, it is not possible to pinpoint the location of the disease locus, especially when a large amount of data is available. With the confidence-set approach, the bottom left histogram in figure 2 shows that the set was never empty, and in only 16 out of 1,000 replications did the set include four markers. Typically, only two markers were included in a confidence set; hence, the disease locus was pinpointed to a 20-cM region, with ⩾99% confidence. With 250 SPs, the maximum number of markers in any confidence set dropped to two, and that occurred only 15% of the time (fig. 2, *bottom right*). These results indicate that, typically, the disease gene was inferred to be within a 10-cM region.

Coverage Probability

We investigated how increased coverage probability may affect the number of markers contained in a confidence set. Each data set consisted of 250 PU SPs. With coverage probability controlled to be 95%, all the confidence sets included marker 15 (fig. 3). Marker 16, the marker next closest to the disease but >5 cM away, was included ~10% of the time. There were also a few instances in which the confidence sets included marker 14. As the coverage probability increased, more confidence sets contained two markers (up to >60% when the coverage probability was ⩾99.99%). Some even contained three markers; however, even with 99.99% coverage probability, only 12% of the confidence sets contained three markers. Furthermore, none of the markers other than the three that were closest to the disease were included in any confidence set, even when the coverage probability was nearly 100%. These results indicate that increasing coverage probability for a given data set increases the number of markers included in the confidence set, as expected; however, even with coverage probability as high as 99.99%, the set is still small enough that inferences about the disease-gene location can be made rather precisely.

Map Density

The effects of marker spacing are presented in table 3. The disease locus was one-third of the way to marker 16 from marker 15. Three different marker densities—1 cM, 5 cM, and 10 cM—were studied for their effects on confidence sets and on the conclusions drawn from the confidence sets. For the PK data, the confidence sets included marker 15, m^*, >99% of the time, for all three levels of map density. Although not all confidence sets included marker 15, all confidence sets included at least one of the three markers closest to the disease. However, as marker spacing increases, the frequency with which the confidence set includes markers other than m^* decreases; for example, marker 14 (i.e., $m^* - 1$) was included in >50% of the confidence sets when the markers were separated by 1 cM but was included in only 1 of the 1,000 confidence sets when the density of the markers was increased to 10 cM; in fact, at a 10-cM density, none of the markers other than the three closest to the

Table 3

Marker-Spacing Effects on Confidence Sets

Inference on Confidence Set[a]	Estimated Probability When Marker Spacing Is					
	1 cM	5 cM	10 cM	1 cM	5 cM	10 cM
	PK Data			PU Data		
$P(m^* \in A)$	.998	1.0	1.0	.998	.999	1.0
$P(m^* + 1 \in A)$	.952	.874	.742	.954	.857	.797
$P(m^* - 1 \in A)$	.533	.013	.001	.521	.024	.002
$P(m^* + 2 \in A)$	.285	.003		.284	.002	
$P(m^* - 2 \in A)$	.059			.072		
$P(m^* + 3 \in A)$	.030			.030		
$P(m^* - 3 \in A)$	.004			.006		
$P(m^* + 4 \in A)$				.002		
$P(\tilde{m} \in A)$	1.0	1.0	1.0	1.0	1.0	1.0
max [#(A)][b]	6	4	3	5	3	3
$P[\#(A) = 1]$	.020	.125	.258	.013	.138	.202
$P[\#(A) = 2]$	.295	.861	.741	.316	.842	.797
$P[\#(A) = 3]$	.503	.013	.001	.484	.020	.001
$P[\#(A) = 4]$	.169	.001		.165		
$P[\#(A) = 5]$	.012			.022		
$P[\#(A) = 6]$	.001					
E[#(A)][c]	2.861	1.890	1.733	2.867	1.882	1.799

[a] All probabilities are estimated by relative frequencies, which are based on 1,000 replications. m^* = marker 15, the marker closest to the disease locus; $m^* + 1$ = the next closest, followed by $m^* - 1$, etc; $\tilde{m}$ = either m^*, $m^* - 1$, or $m^* + 1$.

[b] max[#(A)] = maximum number of markers inferred to be tightly linked.

[c] E[#(A)] = average number of markers inferred to be tightly linked.

disease locus were included in any confidence set, a fact reflected by the blank entries in table 3. This makes intuitive sense: as marker spacing increases, markers other than m^* become farther away from the disease, and, thus, the chances that they will be inferred to be tightly linked become smaller. The observations above are based on marginal inferences for each marker. We can make joint inferences for all markers as well. The average number of markers inferred to be tightly linked decreased with increasing marker spacing, from 2.86 for 1 cM to 1.73 for 10 cM. The maximum number of markers inferred to be tightly linked also decreased with spacing. These observations hold true, qualitatively, for the PU data type as well, as one can see in table 3.

The clear pattern of less-dense maps leading to smaller confidence sets may give one a false impression that it is better to use a coarse map to search for disease genes. In general, it is still better to have a dense map even when more markers may be falsely inferred to be tightly linked; for instance, even when six markers were contained in the confidence set when a 1-cM map (the most extreme case in our simulation) was used, the disease was still localized, with 99% confidence, to a small region of 6 cM (6 × 1 cM); on the other hand, with a 10-cM map, although there may be only three markers contained in a confidence set, the region in which one needs to search for the disease gene was 30 cM (3 × 10 cM) long, five times longer than the region implicated when a 1-cM map was used.

Disease-Gene Location

Various disease-gene locations between markers 15 and 16, at 0 cM, 1 cM, 2 cM, 3 cM, 4 cM, or 5 cM from marker 15, were studied to determine their influences on confidence sets using the PK data type, and the results are plotted in figure 4. None of the markers other than 14, 15, or 16 were included in any confidence set, regardless of the disease-gene location. When the disease gene was placed on top of marker 15, most of the confidence sets included only marker 15. Markers 14 and 16 each were included only ~5% of the time, matching the theoretical results in table 1. When the disease was placed farther away from marker 15 but closer to marker 16, marker 16 was included in the confidence sets more often, as expected. Marker 15 was included in all confidence sets, except at one disease location. The exception occurred when the disease was placed exactly halfway between markers 15 and 16; in this case, both markers were exactly 5 cM from the disease. The proportion of times (per 1,000 replications) that markers 15 and 16 were included in the confidence set was only 97.8% and 97.9%, respectively—less than the specified coverage probability—most likely because of the normal approximation and/or Monte Carlo errors in the simulation. In this case, there are two markers that satisfy the null hypothesis of tight linkage, but there is no need for multiplicity adjustment since we are interested only in the inclusion of at least one of the markers, which, in this case, occurred in 100% of the replicates. In fact, none of the markers other than 15 or 16 were included in any confidence set. These results indicate that the location of the disease gene relative to the two flanking markers does not have a great influence on the number of markers contained in a confidence set.

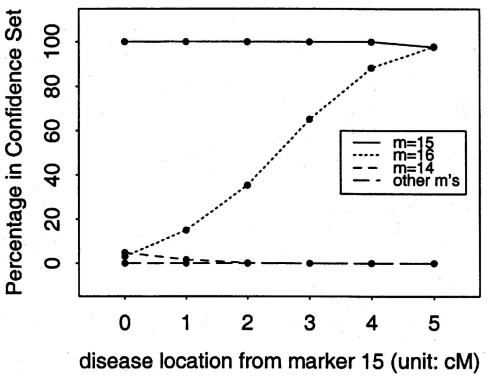

Figure 4 Effects of disease-gene location on the markers included in the confidence set. Each dot represents the relative frequency that the marker was included in the confidence set. Several disease-gene locations, ranging from exactly at marker 15 (0 cM) to exactly halfway between markers 15 and 16 (5 cM), are presented.

Discussion

In the present study, we propose a confidence-set approach for finding tightly linked genomic regions for genome-scan studies. There are two main advantages of this approach. First, we formulate our null hypotheses to correspond to tight linkage to ensure that we need not be concerned with the multiplicity-adjustment problem, which is caused by screening a large number of markers. Second, using this method, one can identify sufficiently localized genomic regions for linkage, so that the need for efforts at further localization is greatly reduced after an initial screen. This approach can thus be viewed as a way of combining an initial scan with fine mapping.

A general principle of multiplicity adjustment is that multiplicity need be adjusted only to the extent that the null hypotheses being tested may be true simultaneously; for example, in average bioequivalence studies, even though two one-sided tests are performed simultaneously, no multiplicity adjustment is needed, because it is impossible to deliver both too much and too little of the

test drug (as compared with the reference drug) at the same time (Berger and Hsu 1996); as another example, in some dose-response studies, step-down testing of efficacy at decreasing doses requires no multiplicity adjustment, because there is, at most, one minimum effective dose (Hsu and Berger 1999). This general principle implies that, when a genetic disorder involves a single disease gene, no multiplicity adjustment is needed, no matter how many markers throughout the genome are tested one by one. The procedure produces a confidence set that includes a marker that is, at most, d_0 from the disease locus with probability p (usually controlled to be $\geqslant$99%). In the description of the method, for ease of presentation, we assume that there exists in the genome a unique marker that is located $\leqslant d_0$ from the disease locus. This assumption is not necessary, as is demonstrated in one of the simulations in which the disease locus was placed exactly halfway between two markers. The probability that at least one of the markers located within distance d_0 is included in the confidence set is greater than the probability that any one of these markers is included in the set, which is controlled to be at least p. A confidence set for the location of the disease gene can also be constructed directly (Rogers et al. 2001).

Now, suppose that the manifestation of a disease involves g genes spread throughout the genome, where some disease genes may be linked to one another. Assume that we have a map of markers covering the whole genome and that there exists a unique marker m_i^* located within d_0 from disease gene i, $i = 1, \cdots, g$. The assumption of uniqueness—which is, again, not necessary, as discussed in the previous paragraph—is used for ease of presentation. Suppose that A is the confidence set to be constructed, such that the probability that A includes all $m_i^*, i = 1, \cdots, g$, is at least $p\ (= 1 - \beta)$. Further suppose that we set the probability of false rejection of a tight linkage to be β^*. Then the genomewide coverage probability of all g genes is at least $1 - g\beta^*$. Hence we can set $\beta^* = \beta/g$ to achieve a genomewide coverage probability of at least $100(1 - \beta)$%, a Bonferroni-like argument. This is, of course, a very crude approximation, especially when the disease genes are linked; thus, more-sophisticated statistical methods need to be developed. Nonetheless, the simple argument above shows that, although we may be screening the genome for thousands of markers when trying to detect signals for linkage, we do not need to adjust for thousands of tests in order to avoid too many false declarations of linkage. We only need to adjust the confidence level (or type I error, in our formulation of the tests) for the number of genes that may be involved in a disease; for example, if 10 genes are assumed to be involved in the manifestation of a disease, then a genomewide coverage probability of 99% implies a 99.9% coverage probability for each disease locus. If the number of genes is actually smaller than the initial guess, the procedure will be conservative, in that the confidence set will include more markers than necessary. One possible approach toward estimating the number of genes that may be involved—and, thus, to attaining a better multiplicity adjustment—is through a two-stage procedure: first, the confidence set is constructed according to an initial guess as to the number of genes involved (usually assumed to be higher than what is anticipated—say, 10); then the number of "well-defined" genomic regions that are included in the confidence set is taken to be the number of disease genes, and a new confidence set is constructed. We caution, though, that such a two-stage approach will be anticonservative if some disease genes are closely linked.

Our procedure is designed to find markers that are tightly linked to disease genes. The sizes of the implicated genomic regions depend on factors such as the amount of data available and the density of the marker map. The probability that is being controlled is the coverage probability. To have high confidence that potential disease loci will not be missed, we recommend setting the coverage probability to be $\geqslant$99%, genomewide. The true coverage probability is usually higher than the level that is set, especially if the distance from the marker to the disease is smaller than the maximum allowable distance. Even with such high coverage probability, results from our simulation study indicate that a typical confidence set would include only a very small number (although usually larger than the targeted number) of markers, with a reasonable amount of data. Furthermore, the more data we have, the more precisely our procedure is able to predict the disease-gene location, a good property that the LOD method does not possess. We recognize that a single study with two simple data types does not warrant general conclusions; thus, further studies investigating the properties of the proposed procedure for general genetic data and models are needed, but the results from the study thus far are extremely encouraging. Finally, we note that our method is applicable even when the disease is not genetic. In such cases, an empty set is highly desirable, since one can then infer that the disease is not affected by any gene. With a sufficient sample size, there is a high probability that our confidence set is empty when the disease is not genetic.

Acknowledgments

We thank two anonymous referees for their helpful comments on the first version of this article. This work was supported in part by National Science Foundation grant DMS-9971770 (to S.L.).

Appendix A

Let f_{θ_m} be the probability mass function of $S_m = \sum_{i=1}^{n} \min\{X_i, 2 - X_i\}$. Then,

$$f_{\theta_m}(t) = \binom{n}{t} 2^t(\theta_m - \theta_m^2)^t(1 - 2\theta_m + 2\theta_m^2)^{n-t},\ t = 0,1,\cdots,n\ .$$

It follows that, for $0 \leqslant \theta_1 < \theta_2 \leqslant \frac{1}{2}$, the likelihood ratio

$$\frac{f_{\theta_2}(t)}{f_{\theta_1}(t)} = \left(\frac{1 - 2\theta_2 + 2\theta_2^2}{1 - 2\theta_1 + 2\theta_1^2}\right)^n \left[\frac{(\theta_2 - \theta_2^2)(1 - 2\theta_1 + 2\theta_1^2)}{(\theta_1 - \theta_1^2)(1 - 2\theta_2 + 2\theta_2^2)}\right]^t$$

is increasing in t, since the expression in the square brackets is >1. Thus, the family of distributions $\{f_{\theta_m}(t):\theta_m \in [0,1/2]\}$ has a monotone likelihood ratio in t. Lemma 2 of Lehmann (1986, p. 85) then implies that, for any $\theta_m < \theta_0$, $P_{\theta_m}(S_m > t) \leqslant P_{\theta_0}(S_m > t)$. Consequently, $\sup_{\theta_m \leqslant \theta_0} P_{\theta_m}(S_m > t) = P_{\theta_0}(S_m > t)$.

References

Berger RL, Hsu JC (1996) Bioequivalence trials, intersection-union tests, and equivalence confidence sets. Stat Sci 11: 283–315

Bernstein F (1931) Zur Grundlegung der Chromosomentheorie der Vererbung beim Menschen. Z Abst Vererb 57:113–138

Bickel PJ, Doksum KA (1977) Mathematical statistics. Prentice-Hall, Englewood Cliffs, NJ, pp 177–182

Curtis D (1996) Genetic dissection of complex traits. Nat Genet 12:356–358

Drigalenko EI, Elston RC (1997) False discoveries in genome scanning. Genet Epidemiol 14:779–784

Edwards JH, Watt DC (1989) Caution in locating the gene(s) for affective disorder. Psychol Med 19:273–275

Elston RC, Lange K (1975) The prior probability of autosomal linkage. Ann Hum Genet 38:341–350

Fisher RA (1935) The detection of linkage with dominant abnormalities. Ann Eugen 6:187–201

Hsu JC, Berger RL (1999) Stepwise confidence intervals without multiplicity adjustment for dose response and toxicity studies. J Am Stat Assoc 94:468–482

Kidd KK, Ott J (1984) Power and sample size in linkage studies: Human Gene Mapping 7 (1984): Seventh International Workshop on Human Gene Mapping. Cytogenet Cell Genet 37:510–511

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

Lange K, Boehnke M (1982) How many polymorphic genes will it need to span the human genome? Am J Hum Genet 34:842–845

Lehmann E (1986) Testing statistical hypotheses. John Wiley and Sons, New York.

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

Ott J (1985) Analysis of human genetic linkage. Johns Hopkins University Press, Baltimore

——— (1999) Analysis of human genetic linkage, 3d ed. Johns Hopkins University Press, Baltimore

Risch N (1991) A note on multiple testing procedures in linkage analysis. Am J Hum Genet 48:1058–1064

Rogers JA, Lin S, Hsu JC (2001) Exact confidence sets for the map location of a disease gene. Technical rep 668. Department of Statistics, Ohio State University, Columbus

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

Thompson EA (1984) Interpretation of LOD scores with a set of marker loci. Genet Epidemiol 1:357–362

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

Zhao LP, Prentice R, Shen F, Hsu L (1999) On the assessment of statistical significance in disease-gene discovery. Am J Hum Genet 64:1739–1753