Genetic Epidemiology 21:31–39 (2001)

Lod Scores for Gene Mapping in the Presence of Marker Map Uncertainty

Heather M. Stringham and Michael Boehnke*

Department of Biostatistics, University of Michigan, Ann Arbor

Multipoint lod scores are typically calculated for a grid of locus positions, moving the putative disease locus across a fixed map of genetic markers. Changing the order of a set of markers and/or the distances between the markers can make a substantial difference in the resulting lod score curve and the location and height of its maximum. The typical approach of using the best maximum likelihood marker map is not easily justified if other marker orders are nearly as likely and give substantially different lod score curves. To deal with this problem, we propose three weighted multipoint lod score statistics that make use of information from all plausible marker orders. In each of these statistics, the information conditional on a particular marker order is included in a weighted sum, with weight equal to the posterior probability of that order. We evaluate the type 1 error rate and power of these three statistics on the basis of results from simulated data, and compare these results to those obtained using the best maximum likelihood map and the map with the true marker order. We find that the lod score based on a weighted sum of maximum likelihoods improves on using only the best maximum likelihood map, having a type 1 error rate and power closest to that of using the true marker order in the simulation scenarios we considered. Genet. Epidemiol. 21:31–39, 2001.

Key words: disease mapping; linkage analysis; linkage maps; marker maps

INTRODUCTION

Multipoint lod scores are typically calculated for a grid of locus positions, moving the putative disease locus across a fixed map of genetic markers. The resulting

Contract grant sponsor: National Institutes of Health; Contract grant numbers: HG00376, T32 HG00040.

*Correspondence to: Michael Boehnke, Ph.D., Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

Received for publication 27 March 2000; Accepted 9 May 2000

multipoint lod score curve is then plotted against this map and the maximum lod score is noted. Particularly given densely-spaced markers, there may not be a single clearly best marker order, and the evidence for linkage may differ substantially depending on the order of a set of markers and/or the distances between the markers. If the best order (by a particular criterion, usually maximum likelihood) is enough better than the second best order, the usual approach is simply to condition analysis on the best order. If there are two orders that are nearly equally likely, one might do the analysis twice, once with each order, and note any important differences in the results. If there are no important differences, the uncertainty in the order may not pose much of a problem. If, however, the two orders yield substantially different results, the uncertainty becomes a concern. If there are more than two marker maps under consideration, the situation becomes even more complicated.

Hanis et al. [1996] described a genome scan for genes contributing to type 2 diabetes susceptibility and presented an example that illustrates this problem. Their data consisted of 330 Mexican-American affected sib pairs from 170 sibships with no parental data. Their best result occurred on chromosome 2 at the tip of the q arm. For this 2qter region, they considered the maps shown in Fig. 1. Their maximum multipoint lod scores ranged from 2.7 to 4.3, depending on which marker order and distances they used.

In this article, we propose a general method to deal with uncertainty in the marker map that acknowledges marker order uncertainty and still provides a single multipoint lod score to assess the evidence for linkage. We calculate three weighted multipoint lod score statistics and evaluate their properties on the basis of results from simulated

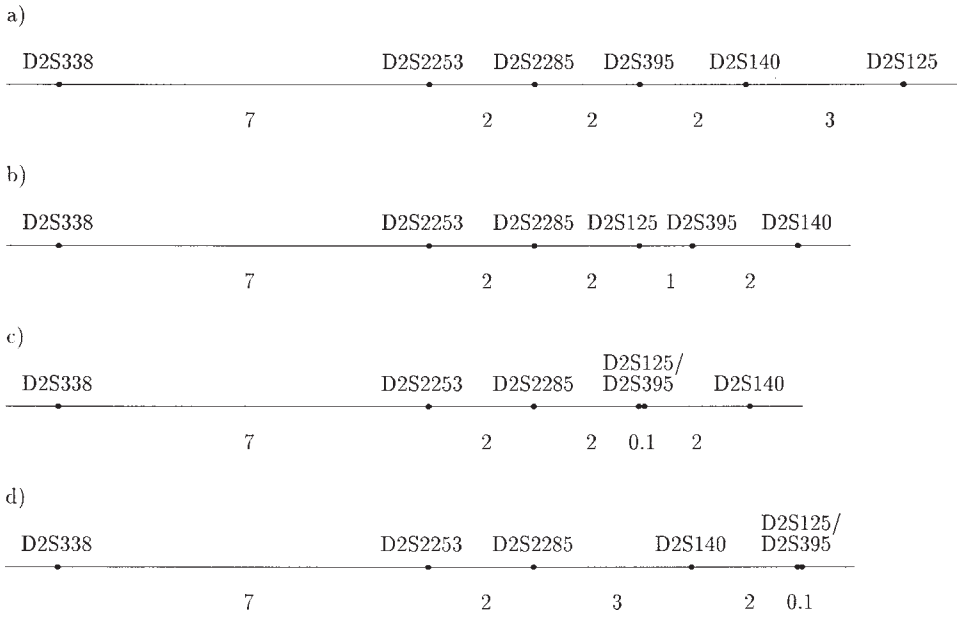

Fig. 1. Plausible maps for a portion of chromosome 2 from Hanis et al. [1996] (N.J. Cox, personal communication), based on maps from (**a**) the Marshfield Clinic database, (**b**) Généthon, (**c,d**) Hanis et al. [1996] data. Distances between markers are given in cM.

data, considering the type 1 error and power. The first statistic is a lod score formed from weighted likelihoods. The second is a weighted average of the lod scores themselves. For the third statistic, we calculate a weighted significance level and then back-transform to its lod score equivalent. We compare the results for these statistics to those obtained using the best maximum likelihood marker map and the map with the true marker order. We find that the lod score formed from weighted likelihoods improves on using the best maximum likelihood map, giving lod scores closest to those obtained using the true marker order in the simulation scenarios we consider.

METHODS

To most accurately localize a disease gene and refine its position on a particular chromosome, ideally, we would jointly estimate disease gene and marker locations. However, we often do not have the appropriate data to do so and are additionally limited by available computing resources, both hardware and software. More realistically, if the positions of a subset of markers are uncertain, we might calculate multipoint lod scores, conditioning on a framework map and integrating over all possible locations for each of the nonframework markers and the disease locus. This is, of course, only nearly ideal since we do not know the locations of the framework markers for certain, and it too would be impractical to implement. Instead, we have chosen to form a weighted sum of the information conditional on the most plausible comprehensive marker orders. We use as weights the posterior probabilities P_γ for the various orders γ, and estimate map distances for each order by maximum likelihood. This approach is computationally simple, and still has the advantage of acknowledging uncertain marker order.

Posterior Probabilities for Marker Orders

The posterior probabilities P_γ can be calculated in a number of different ways [Rogatko and Zacks, 1993; Lange et al., 1995]. One alternative for P_γ is

$$P_\gamma = \frac{\text{Prior}(\gamma)L(\gamma)}{\sum_j \text{Prior}(j)L(j)}. \quad (1)$$

Here, the posterior likelihood for order γ is divided by the sum of the posterior likelihoods for all orders under consideration so that P_γ represents the fraction of the total posterior likelihood attributable to order γ. Since a constant prior is reasonable, such as $2/m!$ where m is the number of markers, equation (1) is equivalent to

$$P_\gamma = \frac{L(\gamma)}{\sum_j L(j)}. \quad (2)$$

For either equation (1) or (2), we could calculate the likelihood $L(\gamma)$ as

$$\int_{\underline{\theta}_\gamma} L(\underline{\theta}_\gamma; \gamma, data)\text{Prior}(\underline{\theta}_\gamma; \gamma) d\underline{\theta}_\gamma,$$

averaging the posterior likelihood for the marker map over $\underline{\theta}_\gamma$. Here, $\underline{\theta}_\gamma$ is the vector of recombination fractions between adjacent markers in order γ with no disease locus included. As another option for $L(\gamma)$, we could substitute $\hat{L}(\gamma) = L(\underline{\hat{\theta}}_\gamma;\gamma,data)$. For this alternative, $\underline{\hat{\theta}}_\gamma$ represents the marker map with all markers fixed at their maximum likelihood estimates (MLEs) for order γ. We use

$$P_\gamma = \frac{\hat{L}(\gamma)}{\sum_j \hat{L}(j)}$$

as it is easier to compute given computations routinely done in marker map construction and can be expected to provide a good approximation to more complicated formulas [Rogatko and Zacks, 1993; Lange et al., 1995]. When the number of possible orders $m!/2$ is large, we approximate the sum in the denominator by including only the most plausible orders j, which could be determined, for example, by eliminating any order with maximum likelihood more than 10,000 times less than that of the most likely order.

Weighted Lod Score Statistics

To take marker map uncertainty into account, we calculate three different weighted multipoint lod score statistics, weighting each order by its posterior probability. First, we calculate

$$\hat{Z}_{wt\ like} = \log_{10} \frac{\sum_\gamma P_\gamma L(\hat{\theta};\gamma,data)}{\sum_\gamma P_\gamma L(\frac{1}{2};\gamma,data)},$$

weighting the likelihood $L(\theta;\gamma,data)$ under each comprehensive marker order γ by its posterior probability P_γ, and summing over all plausible orders γ. Here we define $\hat{\theta}$ to be the MLE of the recombination fraction between the maker map and the disease locus, fixing the map distances for each order γ at their MLEs. We would expect this to be the best approach since it corresponds to comparing the likelihood of the data under the alternative and null hypotheses of linkage and no linkage in a similar manner to the usual unweighted lod score.

Since the likelihoods $L(\theta;\gamma,data)$ may not be obtained from many commonly-used software packages without modification of the code, we calculate, as a second alternative, the weighted average of the maximum lod scores:

$$\hat{Z}_{wt\ lod} = \sum_\gamma P_\gamma \log_{10} \frac{L(\hat{\theta};\gamma,data)}{L(\frac{1}{2};\gamma,data)}.$$

As a third alternative, we calculate weighted significance levels, averaging the *P* values associated with the maximum lod scores for each of the plausible orders, and

back-transforming this weighted-average *P* value to its lod score equivalent, $\hat{Z}_{wt\ pvalue}$. This statistic also does not require the likelihoods to be obtained and should be less susceptible to being dominated by the most likely order than $\hat{Z}_{wt\ lod}$.

Calculating Weighted Lod Scores for Affected Sib Pairs

To calculate the weighted multipoint lod scores for an affected sib pair data set, we make use of the programs CRI-MAP [Lander and Green, 1987] (Green et al., unpublished documentation), to construct multilocus linkage maps, and SIBLINK [Hauser and Boehnke, 1998], to calculate multipoint lod scores based on allele sharing between sib pairs. These lod scores are parameterized by the recombination fraction θ and the vector of allele-sharing probabilities $\mathbf{z} = (z_0, z_1, z_2)$, where z_i is the probability that a sib pair shares i alleles identical by descent (IBD) at the putative disease locus. Assuming a model where risk alleles at the disease locus act additively, $z_0 < ¼$, $z_1 = ½$, and $z_2 > ¼$.

To analyze a particular data set, we first build a framework map of markers using CRI-MAP *build*, adding markers sequentially and eliminating any order whose maximum likelihood is more than 1,000 times less than that of the best map so far. This step reduces the number of comprehensive orders that must be considered in the next step. Second, using CRI-MAP *all*, we generate a list of all comprehensive orders consistent with the framework order that have maximum likelihoods no more than 10,000 times less than the most likely order. We calculate the posterior probabilities P_γ using the maximum likelihoods for each of these orders. We estimate the marker map distances for the comprehensive orders by maximum likelihood using CRI-MAP *fixed*, and calculate multipoint lod scores under each of the orders using SIBLINK. Finally, we calculate the weighted lod scores, using the posterior probabilities P_γ and the maximum lod scores or likelihoods from SIBLINK for each order.

Simulations

Properties of interest for each of the proposed lod scores include the type 1 error rate and the power to detect the presence of a disease gene. We examined these properties by computer simulation, comparing them among our proposed lod scores and with the standard approach of just using the maximum lod score for the best maximum likelihood map. We also compared the results to those obtained using the marker map with map distances estimated by maximum likelihood, conditional on the true order of the markers, which we refer to as the map with the true marker order.

We chose our simulation parameters to mimic the scenario in the Hanis et al. [1996] example. We used the map shown in Fig. 1a, with markers at 0, 7, 9, 11, 13, and 16 cM and disease locus at 15 cM, and simulated four-person nuclear families with two genotyped affected sibs. Fifty such families were simulated with both parents genotyped and were used to estimate the marker maps. To assess evidence for linkage, we simulated 500 affected sib pair families with no parents genotyped. Marker genotypes were generated assuming four equally-frequent alleles at each of the six markers and a random genotyping error rate of 1%. We assumed that the alleles at the disease locus act additively and the recurrence risk ratio λ takes values of 1.0 or 1.4, corresponding to $\mathbf{z} = (0.25, 0.50, 0.25)$ and $\mathbf{z} = (0.179, 0.500, 0.321)$, respectively.

We assessed type 1 error by analyzing 12,500 simulated data sets, generated assuming no disease locus is present ($\lambda = 1.0$), and counting the number of false

positives in the analyses. Power was evaluated by analyzing data generated assuming a disease locus *is* present ($\lambda = 1.4$) and counting the number of simulations that detect the locus. Simulations were excluded when only one comprehensive order was plausible. In addition, for $\lambda = 1.4$, we included only those 833 simulations in which the difference in maximum lod scores for the two best orders was ≥0.5.

RESULTS

The type 1 error rate is similar for our three weighted lod scores and for the lod score computed using the best maximum likelihood map (Table I). For all four of these statistics, the type 1 error is similar to that obtained when the map with the true marker order is used to compute the lod score curve. $\hat{Z}_{wt\ like}$, the lod score formed from weighted likelihoods, has type 1 error closest to that using the map with the true marker order.

The lod score $\hat{Z}_{wt\ like}$ is also closest to using the true marker order in terms of power. As seen in Table II, using the true order gives the highest power of all the lod score statistics, although differences in power are not large. Further, $\hat{Z}_{wt\ like}$ gives the lod score closest on average, and with the least variability, to what we would obtain if we knew the true order of the markers (Table III). The mean difference in lod score from the true order is 0.003 and standard deviation (s.d.) = 0.15 (Table III, columns 1–3). In contrast, using the best maximum likelihood map gives a mean difference in lod score of –0.19 with s.d. = 0.48. $\hat{Z}_{wt\ pvalue}$, the lod score formed from weighted significance levels, gives the largest difference from the true order, and $\hat{Z}_{wt\ lod}$, the weighted maximum lod score, gives values in between those of $\hat{Z}_{wt\ pvalue}$ and $\hat{Z}_{wt\ like}$.

Since the best maximum likelihood map is the same as the map with the true marker order when the best map orders the markers correctly, it is useful to look at the case where the best order is not the true order. When the best maximum likelihood order is correct, the power and average lod scores are nearly identical for $\hat{Z}_{wt\ like}$ and the lod score for the best order (Tables II and III, columns 4–6). When the best order is incorrect, $\hat{Z}_{wt\ like}$ has power closest to that for the true marker order with mean difference in lod score from the true order of 0.004 and s.d. = 0.19 (Tables II and III, columns 7–9). In contrast, the lod score using the best maximum likelihood map has a mean difference in lod score of –0.34 with s.d. = 0.61 when the best map does not order the markers correctly. We observe similar trends in simulations based on different scenarios such as maps with markers equally spaced at 5 cM (data not shown).

TABLE I. Type 1 Error Rate

$\lambda = 1.0$	N = 12,500			Best order correct (N = 6,755)			Best order wrong (N = 5,745)		
	Fraction of lod scores			Fraction of lod scores			Fraction of lod scores		
$\hat{Z}$	≥3.0	≥2.0	≥1.0	≥3.0	≥2.0	≥1.0	≥3.0	≥2.0	≥1.0
True order	0.0002	0.0035	0.0357	0.0004	0.0025	0.0330	0.0000	0.0047	0.0388
Best map	0.0002	0.0025	0.0324	0.0004	0.0025	0.0330	0.0000	0.0024	0.0317
Wt like	0.0002	0.0034	0.0353	0.0004	0.0025	0.0333	0.0000	0.0044	0.0376
Wt lod	0.0001	0.0026	0.0305	0.0001	0.0025	0.0306	0.0000	0.0026	0.0303
Wt *pvalue*	0.0001	0.0026	0.0298	0.0001	0.0025	0.0302	0.0000	0.0026	0.0294

TABLE II. Power

$\lambda = 1.4$	N = 833			Best order correct (N = 377)			Best order wrong (N = 456)		
	Fraction of lod scores			Fraction of lod scores			Fraction of lod scores		
$\hat{Z}$	≥3.0	≥2.0	≥1.0	≥3.0	≥2.0	≥1.0	≥3.0	≥2.0	≥1.0
True order	0.84	0.97	1.00	0.84	0.97	1.00	0.84	0.97	1.00
Best map	0.80	0.95	1.00	0.84	0.97	1.00	0.77	0.93	1.00
Wt like	0.84	0.97	1.00	0.84	0.97	1.00	0.83	0.97	1.00
Wt lod	0.81	0.96	1.00	0.83	0.97	1.00	0.79	0.95	1.00
Wt *pvalue*	0.79	0.95	1.00	0.82	0.96	1.00	0.77	0.94	1.00

DISCUSSION

We have developed a general method and three multipoint lod scores for use when there is uncertainty in the marker map. The three lod scores take uncertainty in marker order into account and are easy to compute with quantities available from existing software. In the simulation scenarios we considered, type 1 error and power for the weighted-likelihood lod score, $\hat{Z}_{wt\ like}$, are closest to those for the map with the true marker order and, therefore, improve on the simpler and more typical approach of using the best maximum likelihood map. Although we have implemented our method using the program SIBLINK to calculate the multipoint lod scores for affected sib pair data, the method is not limited to affected sib pair data and could be used with any family data for which multipoint likelihoods and lod scores can be calculated.

We found it surprisingly difficult to duplicate by simulation the examples we have seen in the literature where choice of map substantially affects the lod score curve. These examples often utilize maps from different sources (such as Généthon, Marshfield, or CEPH) and we have found that maps generated from different sources typically result in lod score curves that are more different from each other than maps generated from the same data set. Maps from different sources cannot be ranked and posterior probabilities cannot be calculated, so that our method cannot easily be applied in this case. In light of this, we would recommend using the best available single data set to estimate marker maps whenever possible. However, when this cannot be done, it may be possible to obtain likelihoods for marker orders obtained from different sources by calculating the likelihoods for these orders using

TABLE III. Average Maximum Lod Score and Mean Difference from Maximum Lod Score Using Map With True Marker Order (ΔLod)

$\lambda = 1.4$	N = 833			Best order correct (N = 377)			Best order wrong (N = 456)		
	Fraction of lod scores			Fraction of lod scores			Fraction of lod scores		
$\hat{Z}$	Lod	ΔLod	s.d.	Lod	ΔLod	s.d.	Lod	ΔLod	s.d.
True order	4.74	0	0	4.78	0	0	4.70	0	0
Best map	4.55	–0.19	0.48	4.78	0	0	4.36	–0.34	0.61
Wt like	4.74	0.00	0.15	4.79	0.00	0.07	4.71	0.00	0.19
Wt lod	4.54	–0.19	0.38	4.71	–0.08	0.19	4.41	–0.29	0.46
Wt *pvalue*	4.43	–0.31	0.40	4.61	–0.18	0.27	4.28	–0.42	0.46

the CEPH families. Ranking orders by use of a data set that did not generate the orders is clearly not ideal, but it can often be quickly and easily accomplished using, for example, the web-based server MAP-O-MAT [Matise and Gitlin, 1999]. Problems still remain, however, if markers of interest are not typed or available for the CEPH families or if the different marker maps include different subsets of markers. Note that other ordering methods, such as radiation hybrid mapping, could be used to generate the list of plausible orders, as long as a likelihood for each order can be produced.

A disadvantage of our method is that it does not provide an estimate of location for the disease locus. We have modified the method to calculate the weighted lod scores for specific values of the recombination fraction between a fixed marker framework map and the disease locus, instead of using the MLE $\hat{\theta}$ for each order. After calculating these lod scores for a grid of recombination fractions, we can plot the lod scores against the framework map and note the maximum lod score, along with the location at which it occurs, in the usual manner. This provides a location estimate for the disease locus relative to a framework map of markers. To implement this modification, marker map distances for the plausible comprehensive orders must be estimated, fixing the positions of a framework of markers. Type 1 error and power are similar to that of our original method, but we have not found an improvement in bias of the location estimate over using the best maximum likelihood map (data not shown).

For data generated assuming a disease locus is present ($\lambda = 1.4$), we chose to consider only those simulations for which the difference in lod scores for the best maximum likelihood map and the second best map was ≥0.5. By so doing, we hoped to concentrate on cases in which map uncertainty had the greatest effect on the lod score curve and thus would be of greatest concern. If we remove this restriction and look at simulations without regard to difference in lod scores, we see similar trends for the scenarios we considered. The power for $\hat{Z}_{wt\ like}$ is closest to that for the lod score using the true marker order. Further, $\hat{Z}_{wt\ like}$ is closest on average, with least variability, to the lod score for the best order, improving on the maximum lod score for the best map when this map does not order the markers correctly (data not shown). Although the magnitude of the difference between using $\hat{Z}_{wt\ like}$ and using the best maximum likelihood map is greater for the simulations where uncertainty is of greatest concern, $\hat{Z}_{wt\ like}$ improves on using the best maximum likelihood map regardless of the degree of map uncertainty. In light of these results, the multipoint lod score, $\hat{Z}_{wt\ like}$, should provide a useful tool for gene mapping whenever map uncertainty is of concern.

ELECTRONIC-DATABASE INFORMATION

URLs for data in this article are as follows:

CEPH Genotype database, http://www.cephb.fr/cephdb/
CRI-MAP, http://linkage.rockefeller.edu/multimap/crimap/
CRI-MAP documentation, http://linkage.rockefeller.edu/soft/crimap/
Généthon, http://www.genethon.fr/
MAP-O-MAT, http://compgen.rutgers.edu/mapomat/
Marshfield, http://research.marshfieldclinic.org/genetics/

ACKNOWLEDGMENTS

Support for this work was provided by research grant HG00376 (M.B.) and training grant T32 HG00040 (H.M.S.) from the National Institutes of Health. We thank William L. Duren for his help in modifying the CRI-MAP code to allow for fixed distances between non-adjacent markers.

REFERENCES

Hanis CL, Boerwinkle E, Chakraborty R, Ellsworth DL, Concannon P, Stirling B, Morrison VA, et al. 1996. A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. Nat Genet 13:161–6.

Hauser ER, Boehnke M. 1998. Genetic linkage analysis of complex genetic traits by using affected sibling pairs. Biometrics 54:1238–46.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci USA 84:2363–7.

Lange K, Boehnke M, Cox DR, Lunetta KL. 1995. Statistical methods for polyploid radiation hybrid mapping. Genome Res 5:136–50.

Matise TC, Gitlin JA. 1999. MAP-O-MAT: marker-based linkage mapping on the World Wide Web. Am J Hum Genet Suppl 65:A435.

Rogatko A, Zacks S. 1993. Ordering genes: controlling the decision-error probabilities. Am J Hum Genet 52:947–57.