Genetic Epidemiology 9:45–59 (1992)

Adequacy of Single-Locus Approximations for Linkage Analyses of Oligogenic Traits

Veronica J. Vieland, Susan E. Hodge, and David A. Greenberg

Departments of Biostatistics and Psychiatry, Columbia University and the New York State Psychiatric Institute (V.J.V., S.E.H.), and Department of Psychiatry, Mount Sinai Medical Center (D.A.G.), New York, New York

When a disease is controlled by two or more mendelian loci acting epistatically, it can be modeled in a linkage analysis as a single-locus mendelian disease with reduced penetrance. However, the reliability of such an approximation has not yet been demonstrated. This study evaluates the adequacy of such single-locus approximations, when the disease under investigation is determined by two loci, one of which is tightly linked to a genetic marker.

A wide range of two-locus models were simulated, and analyzed under both the correct two-locus model and under a single-locus approximation to that model. In general, the single-locus approximations yielded lod scores very close to the correct ones, but estimates of θ tended to be upwardly biased. We conclude that a single-locus linkage analysis will, in general, provide an excellent approximation to a correct (two-locus) linkage analysis of epistatic two-locus diseases. This enables researchers to continue to use single-locus linkage analyses when two-locus disease transmission is a possibility, and it validates linkage findings already obtained under single-locus analysis, even if the disease under investigation proves ultimately to be governed by two mendelian loci. We also examine alternative methods for obtaining parameter estimates for the single-locus approximations, and we discuss both generalizations and limitations of our findings.

Key words: Two-locus models, epistasis, reduced penetrance, simulation

INTRODUCTION

When a disease is controlled by two or more mendelian loci acting epistatically, it is possible to model the disease in a linkage analysis as a single-locus (SL) mendelian trait with reduced penetrance. The use of SL modeling in this case constitutes an

Received for publication September 26, 1991; revision accepted January 13, 1992.

Address reprint requests to Veronica Vieland, New York State Psychiatric Institute, Box 14, 722 West 168th Street, New York, NY 10032.

approximation to a correct (oligogenic) linkage analysis. While there have been indications that SL linkage models may be reliably employed in the analysis of oligogenic traits (see below), the accuracy of the SL approximation has not previously been rigorously established.

A preliminary indication of the accuracy of such SL approximations came from a simulation study of two-locus (2L) genetic models. One can think of a 2L disease as analogous to a SL disease with a form of reduced penetrance caused by the action of a second gene. Greenberg and Hodge [1989] investigated the impact of such genetically caused "reduced penetrance" on SL linkage analyses. They used SL models with random reduced penetrance in the analysis of two types of linkage data: (i) data simulated under a SL generating model with reduced penetrance; and (ii) data simulated under a 2L generating model with complete penetrance (i.e., no additional reduction in penetrance beyond the action of the second gene). Their results suggested that the cause of the reduced penetrance ("random" vs. the action of a second gene) would not exert substantial influence on the results of a SL linkage analysis. Additional work [Greenberg, 1990] corroborated this result. However, neither of these studies undertook a direct assessment of the *accuracy* of SL linkage models in analyzing diseases governed by two loci, since they did not have the software for a 2L linkage analysis.

More recently [Vieland et al., 1992], we extended the original Greenberg and Hodge work in order to *directly* assess the accuracy of SL linkage analyses for 2L diseases. We simulated data using the same 2L models that Greenberg and Hodge had used, but analyzed them both under SL analysis models and under the correct 2L analysis models. This was the first study to use a two-locus *analysis* model, and therefore the first study to determine the *accuracy* of a SL approximation to a (correct) 2L linkage analysis. For the limited set of models examined, the SL analyses provided an excellent approximation to the (correct) 2L analyses. This was further evidence that researchers could rely upon SL linkage analysis in investigations of 2L diseases. However, that study was limited in that it considered only a small number of generating models, and employed moderate sample sizes. Thus, the results could not be readily generalized.

The current study broadens and generalizes the conclusions reached in Vieland et al., viz., that a simple SL approximation will perform well in linkage analyses of traits governed by two loci. In the present study, we systematically examine a broad range of models; we use larger datasets in the simulations; we examine sampling variability; and we compare and evaluate alternative SL approximations to a 2L model. (However, we simulate only nuclear families, see also **Restrictions on generalizability** in the Discussion.) As a result, we can now generalize the earlier findings with confidence: SL analysis models are appropriate for conducting linkage analyses for a broad class of models of epistatic oligogenic transmission.

MODELS AND METHODS

We simulated nuclear family data under 2L *generating models,* and then analyzed the data under: (a) the correct *2L analysis model* (i.e., a model which is identical to the one used to generate the data), and (b) a "corresponding" *SL analysis model*. In this section, we describe the generating models and the analysis models, along with their relationships to one another.

I. Generating Models

Data were generated under models in which affectedness is fully determined by two interacting loci, each of which shows mendelian inheritance. The two disease loci are assumed to be unlinked to one another, and one of the two is linked to a marker. The models thus fall into four classes: those in which the mode of inheritance at both disease loci is recessive (RR models); those in which the mode of inheritance at both loci is dominant (DD models); those in which the mode of inheritance at the linked locus is recessive, and the mode of inheritance at the unlinked locus is dominant (RD models); and those in which the mode of inheritance at the linked locus is dominant, and at the unlinked, recessive (DR models).

Given the mode of inheritance at each locus, each generating model is defined by two parameters, viz., the disease allele frequency at each locus (q_1, q_2). (A parameter for additional reduced penetrance could be added to the model, but will not be considered here. See the Discussion for comments on this point.) This genetic model gives rise to two observable quantities:

$$\text{(i) } K \equiv \text{disease prevalence}$$

$$\text{(ii) } \Psi \equiv \text{population segregation ratio} \equiv \frac{\textit{prevalence of disease}}{\textit{prevalence of segregating mating types}}.$$

Details of the 2L generating models are given in Appendix A.

Table I shows the parameters for the 18 different generating models used in this study, grouped by mode of inheritance (RR, DD, etc.) and by prevalence.

TABLE I. 2L Generating Models*

Model	q_1	q_2	K	Ψ
RR(1)	0.75	0.03	0.0005	0.16
RR(2)	0.15	0.15	0.0005	0.08
RR(3)	0.03	0.75	0.0005	0.16
RR(4)	0.77	0.29	0.05	0.23
RR(5)	0.47	0.48	0.05	0.18
RR(6)	0.29	0.77	0.05	0.23
DD(1)	0.20	0.01	0.01	0.31
DD(2)	0.05	0.05	0.01	0.28
DD(3)	0.01	0.20	0.01	0.31
DD(4)	0.70	0.06	0.1	0.48
DD(5)	0.18	0.17	0.1	0.35
DD(6)	0.06	0.70	0.1	0.48
RD(1)	0.55	0.02	0.01	0.24
RD(2)	0.17	0.19	0.01	0.18
RD(3)	0.11	0.58	0.01	0.24
DR(1)	0.58	0.11	0.01	0.24
DR(2)	0.19	0.17	0.01	0.18
DR(3)	0.02	0.55	0.01	0.24

*The first locus is the linked one. Each group of three models represents a single choice of mode of inheritance at each locus, and prevalence.

Within each group, the disease allele frequency at the linked locus is allowed to range from (relatively) high to (relatively) low. For a fixed prevalence, this determines a corresponding range of disease allele frequencies at the unlinked locus going from low to high. The lower the disease allele frequency at the *linked* locus, the *more* informative the model is for linkage (the higher the resulting lod scores will be); the lower the allele frequency at the *unlinked* locus, the *less* informative the model is for linkage (the lower the resulting lod scores will be; this is analogous to a reduction in penetrance). Thus, for each prevalence value, the models become more informative for linkage as one reads down the Table.

II. Simulation Procedures

Nuclear families were simulated under each of the 18 2L generating models with a recombination fraction $\theta = 0$ between the first (linked) disease locus and the marker locus, with no differential recombination with respect to sex.

Only fully informative matings (AB × CD) were generated at the marker locus. Family sizes were determined according to a negative binomial distribution (mean = 2.8, $\sigma = 2.3$) [Cavalli-Sforza and Bodmer, 1971; see also Greenberg, 1984 for additional details of the simulation procedures]. Standard ascertainment assumptions were built into the program, and the ascertainment probability was 0.05.

III. Analysis Models

The simulated datasets were analyzed under two types of analysis models, 2L and SL, which will be described in turn.

(1) 2L analysis models. For each generating model, the simulated data were analyzed under the correct *2L analysis model,* i.e., a model employing the same parameters used to *generate* the data (the mode of inheritance at each locus, and the gene frequencies q_1, q_2).

(2) SL analysis models. The data for each generating model were also analyzed under a *corresponding SL analysis model*. We wanted the SL analysis model to be one which a researcher could derive from realistically available information. Therefore, the SL analysis models we use are derived from the observable quantities K (prevalence) and Ψ (population segregation ratio) of each generating model. Specifically, we define a *corresponding* SL analysis model as a model derived from K and Ψ (for a specific generating model) via the following procedure [following Greenberg and Hodge, 1989]:

First, a SL penetrance f is calculated as $\Psi/.25$, if the linked locus is recessive; or $\Psi/.5$, if the linked locus is dominant.

Gene frequencies are then calculated by the formulas

$$q = \begin{cases} \sqrt{\frac{K}{f}} & ; \textit{recessive disease} \\ \sqrt{1 - \frac{K}{f}} & ; \textit{dominant disease} \end{cases}$$

where q = frequency of a, $(1-q)$ = frequency of A; a = disease allele for a recessive disease, A = disease allele for a dominant disease. Table II shows the parameters used by the SL analysis models.

TABLE II. Corresponding SL Analysis Models*

Model	q	f
RR(1)	0.03	0.64
RR(2)	0.04	0.32
RR(3)	0.03	0.64
RR(4)	0.24	0.90
RR(5)	0.26	0.74
RR(6)	0.24	0.90
DD(1)	0.01	0.62
DD(2)	0.01	0.55
DD(3)	0.01	0.62
DD(4)	0.05	0.97
DD(5)	0.07	0.70
DD(6)	0.05	0.97
RD(1)	0.10	0.97
RD(2)	0.12	0.72
RD(3)	0.10	0.95
DR(1)	0.01	0.48
DR(2)	0.01	0.36
DR(3)	0.01	0.48

*The models in this Table have the same designation (RR(1), RR(2), etc.) as the corresponding 2L models specified in Table I.

Note that once the mode of inheritance is chosen, the procedure for deriving a corresponding SL analysis model makes no distinction between the linked and the unlinked disease loci. Thus, for instance, the generating models RR(1) and RR(3) give rise to the same corresponding SL analysis model. The corresponding SL gene frequency estimate tends to be quite close to the correct (generating) gene frequency for the *linked* locus for the most informative models, but quite different from the correct gene frequency at the linked locus for the least informative models. We will refer to this pattern in the Results and consider its significance in the Discussion.

IV. Analysis Procedures

Two-locus analyses were conducted using a 2L version of LINKAGE, TMLINK [Lathrop and Ott, 1990]. The program has been extensively tested [see Vieland et al., 1992; Boehnke, personal communication], and appears to be accurate to at least the fifth significant digit. Single-locus analyses were also conducted in TMLINK, by setting the disease allele frequency at the unlinked locus equal to unity. Lod scores were calculated at $\theta = 0.0, 0.02, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.3$, and 0.4, and the maximum lod over these values was obtained (without interpolation). One thousand nuclear families were generated for each model. These families were then analyzed as 50 datasets consisting of 20 nuclear families each, with means and standard deviations computed across the 50 datasets.

RESULTS

Table III shows the mean maximum lod score (Z_{MAX}) and the mean estimate of the recombination fraction ($\hat{\theta}$) for each analysis, along with their standard deviations.

TABLE III. 2L and Corresponding SL Results*

	2L Results				SL Results			
Model	$\hat{\theta}$	(s.d.)	Z_{MAX}	(s.d.)	$\hat{\theta}$	(s.d.)	Z_{MAX}	(s.d.)
RR(1)	0.14	(0.20)	0.2	(0.2)	0.32	(0.13)	0.3	(0.4)
RR(2)	0.02	(0.04)	1.5	(0.8)	0.04	(0.05)	1.6	(0.9)
RR(3)	0.00	(0.01)	5.6	(1.6)	0.00	(0.01)	5.4	(1.5)
RR(4)	0.13	(0.20)	0.4	(0.4)	0.30	(0.12)	0.4	(0.5)
RR(5)	0.04	(0.08)	2.1	(1.3)	0.12	(0.09)	2.1	(1.4)
RR(6)	0.01	(0.03)	5.4	(2.2)	0.05	(0.04)	5.1	(2.3)
DD(1)	0.08	(0.15)	1.3	(0.9)	0.18	(0.14)	1.3	(1.1)
DD(2)	0.04	(0.07)	2.3	(1.3)	0.05	(0.07)	2.3	(1.4)
DD(3)	0.01	(0.02)	4.6	(1.8)	0.01	(0.02)	4.5	(1.7)
DD(4)	0.21	(0.23)	0.1	(0.2)	0.42	(0.09)	0.1	(0.3)
DD(5)	0.03	(0.05)	2.4	(1.4)	0.08	(0.08)	2.4	(1.5)
DD(6)	0.01	(0.01)	9.6	(2.5)	0.02	(0.02)	9.3	(2.6)
RD(1)	0.05	(0.09)	1.4	(1.1)	0.27	(0.12)	1.0	(1.1)
RD(2)	0.02	(0.04)	4.1	(1.8)	0.03	(0.05)	4.1	(1.9)
RD(3)	0.00	(0.01)	7.9	(3.0)	0.02	(0.02)	7.7	(3.2)
DR(1)	0.26	(0.25)	0.1	(0.1)	0.35	(0.18)	0.2	(0.5)
DR(2)	0.12	(0.18)	0.6	(0.5)	0.16	(0.17)	0.6	(0.7)
DR(3)	0.02	(0.05)	3.7	(1.9)	0.02	(0.04)	3.2	(1.5)

*See Tables I and II for the analysis models used here.

I. Mean Maximum Lod Scores (Z_{MAX})

The mean maximum lod scores under the 2L and SL analyses are uniformly close to each other. Among the more informative models, the SL analyses underestimate the 2L mean maximum lod scores by only 2–15%.

Some of the generating models fail to provide enough linkage information to permit detection of linkage in reasonably sized datasets of nuclear families. For example, the RR(1) model yields an average lod score of approximately 0.01 per family. These models are also the ones for which the SL analysis parameters appear to give the poorest fit. Yet even for the least informative generating models in each group, there is remarkable agreement between the SL and 2L mean maximum lod scores.

II. Estimates of the Recombination Fraction ($\hat{\theta}$)

Within each group, the 2L estimates of θ converge towards 0 as the disease allele frequency at the linked locus gets lower. The SL estimates of θ do as well. However, relative to the 2L estimates, the SL estimates of θ tend to be upwardly biased, particularly for the less informative generating models. This is not surprising, since the true recombination fraction is 0, and any error must therefore consist in an *over*count of recombinants.

III. Standard Deviations

In general, the standard deviations are large, both for $\hat{\theta}$ and Z_{MAX}, but strikingly similar for the SL and 2L analyses. This confirms that the SL and 2L analyses are providing statistics with very similar distributions. Note that these standard deviations represent variability *within* each analysis (SL or 2L), and not variability of the

pairwise (dataset by dataset) differences *between* the SL and 2L analyses (see also the Discussion).

We chose a sample size of 20 families per dataset because this figure seemed realistic for family studies. To rule out the possibility that discrepancies between the SL and 2L analyses would increase if the amount of linkage information in each dataset were increased, we reanalyzed the linkage results for each generating model as a single dataset of 1,000 families. In these analyses, there was even less discrepancy between the SL and 2L estimates of θ (as both tended towards 0), and agreement on Z_{MAX} between the two types of analysis remained excellent. For the most informative model within each group, the SL analysis underestimated the 2L lod score by only 2–16%.

DISCUSSION

There are several reasons why a SL linkage analysis might be performed for a 2L disease. In some cases, epidemiologic data may appear consistent with SL mendelian transmission, and SL linkage analyses may be performed under the erroneous belief that the model is correct. In other cases, the disease may appear to be compatible with two-locus (2L) transmission, but researchers will have lacked the computer programs to carry out 2L linkage analysis. And in general, SL linkage analysis has the advantage over 2L analysis of familiarity, and (relative) computational simplicity.

However, if the results of SL linkage analyses of 2L traits are to be interpretable, it is essential to establish the accuracy of the SL approximation. This point is relevant not only to the planning of linkage analyses, but to the interpretation of analyses performed in the past as well. For linkage analyses which have already been performed under SL transmission models, or for linkage analyses of complex disorders with unknown or unresolvable modes of transmission, it is important to gauge the validity of results obtained from SL analyses in the event that the disease proves to be determined by two (or even more) loci.

Two-locus transmission represents a special case of oligogenic models which is of particular interest for two reasons. The first is that numerous biologic traits are known to be governed by two loci. Strickberger [1976], for example, lists twelve distinct known patterns of 2L determination of a single (qualitative) trait in various plant and animal species. There are also human diseases for which 2L transmission is suspected, for instance, insulin-dependent diabetes mellitus (type 1) [Thomson, 1980; Todd et al., 1987], or juvenile myoclonic epilepsy [Greenberg et al., 1988].

The second reason is that, for dichotomous traits, a two-locus epistatic model will produce more extreme departures from a SL model with random reduced penetrance than will epistasis involving several loci. The more genes (or other factors) that are involved in modification of a major gene effect, the more their collective contribution will appear as random reduced penetrance. Therefore, 2L models should be the *least* amenable to approximation by simple SL analyses.

The generating models we examined cover a broad range of epistatic 2L genetic models. If the prevalence of a 2L disease is fixed, admissible pairs of gene frequencies for the loci are bounded [Defrise-Gussenhoven, 1962; Greenberg, 1981], thereby restricting the range of admissible 2L models. We examined generating models spanning: (a) all four classes of simple modes of inheritance (RR, DD, RD, DR); (b) realistic prevalence levels; and (c) a wide range of gene frequencies within each group defined by (a)

and (b). Thus, the 18 generating models used in this study span a broad range of the admissible 2L models.

The results of this study confirm unequivocally our starting hypothesis: a "corresponding" SL analysis model, as we have defined it, will yield lod scores which reliably approximate 2L lod scores, for linkage analyses of diseases governed by two epistatic mendelian loci. However, estimates of θ obtained by SL linkage analysis may be quite inaccurate. The accuracy of the SL analyses in approximating 2L lod scores means that investigators can continue to employ SL linkage analyses for the study of diseases which are governed by two mendelian loci, with little or no reduction in power to detect linkage, and with little risk of falsely excluding linkage. It also means that linkage analyses which have already been performed using SL analysis models remain interpretable if the disease under investigation proves to be governed by two mendelian loci. (However, as long as the true genetic model is unknown, all analyses must be interpreted with caution.)

In what follows, we offer some further observations on: (I) choosing the SL analysis model, (II) quantifying the reliability of the SL lod scores, and (III) generalizing these results to genetic models other than those explicitly considered here.

I. Choosing the SL Analysis Model

We chose to define a "corresponding" SL analysis model as we did, because such an analysis model is easily obtainable by any investigator with access to prevalence and segregation ratio data for the disease under investigation. However, the SL models were derived assuming that the mode of inheritance *at the linked locus* can be correctly specified. In this section, we (1) comment on what happens when the mode of inheritance at the linked locus is not correctly specified. We then (2) comment on alternative procedures for fixing the SL model parameters. Finally, we (3) consider whether one can define a SL model *exactly* corresponding to a simple 2L model.

(1) Using the correct mode of inheritance at the linked locus. In general, 2L RR generating models will appear to a SL segregation analysis as recessives with reduced penetrance, and DD models will appear as dominants with reduced penetrance. However, RD and DR models may appear as recessive or dominant depending on the particular configuration of gene frequencies at each locus. With this in mind, Greenberg and Hodge [1989], and then Vieland et al. [1992], evaluated analysis models in which the mode of inheritance at the linked locus was misspecified. Both studies found that this could result in substantial distortion of linkage results. However, their results were not readily generalizable.

In order to confirm their conclusions regarding misspecification of the mode of inheritance in the SL analysis, we analyzed the DR and RD models again, but this time under a single locus model with the mode of inheritance at the linked locus *mis*specified. The parameters of the single locus model (q, f) were calculated under the wrong mode of inheritance for the linked locus, but otherwise according to exactly the same procedure described in the Models and Methods section above. Table IV shows the results.

When the mode of inheritance at the linked locus is misspecified, the mean maximum lod scores show a marked decrease compared to the SL analysis with mode of inheritance correctly specified, and estimates of θ are, in general, higher. This is consistent with the findings of both earlier studies. Thus, if the prevalence and segrega-

TABLE IV. SL Analyses Under the Wrong Mode of Inheritance at the Linked Locus, Compared to SL Analyes Under the Correct Mode of Inheritance*,**

	Wrong mode of inheritance		Correct mode of inheritance	
Model	$\hat{\theta}$	Z_{MAX}	$\hat{\theta}$	Z_{MAX}
RD(1)	0.25	0.5	0.27	1.0
RD(2)	0.06	1.5	0.03	4.1
RD(3)	0.04	2.3	0.02	7.7
DR(1)	0.42	0.1	0.35	0.2
DR(2)	0.33	0.3	0.16	0.6
DR(3)	0.38	0.2	0.02	3.2

*Correct mode of inheritance results repeated from Table III.
**Standard deviations have been omitted from the Table. They are comparable to those found in Table III.

tion ratio of a disorder are consistent with two-locus RD or DR transmission, it may be desirable to analyze the data under both SL recessive and dominant models [see also Greenberg, 1989].

(2) Alternative approximations to 2L models. The approach to deriving corresponding SL models used in this paper has the advantage of simplicity. But by the same token, it is possible that a different approach could improve the performance of the SL analyses. We saw two possible weaknesses in the procedure described in the Models and Methods section. First, it does not allow for a distinction between models yielding the same observable quantities K, Ψ) on the basis of which disease locus is the *linked* locus. Second, it tacitly depends on the assumption of a rare disease. We refined the procedure for deriving the SL models in two ways, each designed to address one of these issues. Each of these refinements will be discussed in turn. We will then discuss the issue of whether an optimal SL analysis model can be found for a 2L generating model.

(a) First refinement: the "2L-based" approximation. If there exists evidence that a disease is in fact governed by two loci, it is possible to accurately fit a 2L model [Greenberg, 1981]. In that case, one can derive a SL approximation *in light of* the (approximately) correct 2L parameters. That is, one can make use of knowledge of the mode of inheritance at each of the loci, and the estimated allele frequencies at each locus. While it is still not possible to determine *which* of the disease loci will be linked to a given marker, it is then possible to fit two alternative models, allowing for linkage to each disease locus in turn.

We defined a "2L-based" approach to deriving the corresponding SL analysis model as follows. Use the (true) gene frequency at the *linked* locus as the SL gene frequency q. Since the penetrance should be the probability that an individual with the disease genotype at the linked locus has the disease phenotype, we take as the SL penetrance parameter f the probability that an individual who is not a founder has the disease genotype at the *unlinked* locus. This quantity is simply the segregation ratio at the unlinked locus. That is, if the mode of inheritance at the unlinked locus is recessive, we use $1/(1 + p_2)^2$ as the SL penetrance, where p_2 is the frequency of the normal allele at that locus; if the mode of inheritance at the unlinked locus is dominant, we use $1/(1 + q_2^2)$ as the SL penetrance, where q_2 is the frequency of the disease allele at that locus. Note that we elected to base the SL estimate of f on the non-founders within the pedigree. Consideration of the founders might lead one to choose a different approach to deriving f. We comment further on this point below.

TABLE V. Results Using the "2L-Based" SL Analysis Model, Compared to Original SL Results*,

	2L-based SL model				Original SL model	
Model	q	f	$\hat{\theta}$	Z_{MAX}	$\hat{\theta}$	Z_{MAX}
RR(1)	0.75	0.26	0.15	0.2	0.32	0.3
RR(2)	0.15	0.29	0.02	1.5	0.04	1.6
RR(4)	0.77	0.34	0.12	0.4	0.30	0.4
RR(5)	0.47	0.43	0.02	1.7	0.12	2.1
RR(6)	0.29	0.66	0.01	4.8	0.05	5.1
DD(1)	0.20	0.51	0.10	1.3	0.18	1.3
DD(4)	0.70	0.53	0.18	0.1	0.42	0.1
RD(1)	0.55	0.50	0.06	1.3	0.27	1.0
DR(1)	0.58	0.28	0.28	0.1	0.35	0.2

*Original SL results repeated from Table III.
**Standard deviations have been omitted from the Table. They are comparable to those found in Table III.

We reanalyzed the data for selected generating models using corresponding SL analysis models derived according to this "2L-based" procedure. Table V shows the SL parameters, and the effects of refining the SL analysis models in this way. The SL lod scores for the original corresponding SL analysis models were so close to the 2L lod scores, that there is little room for improvement in Z_{MAX}. However, there is considerable room for improvement in $\hat{\theta}$. The 2L-based SL analyses show either little improvement in Z_{max}, or, in many cases, slight decrements over the crude SL approximations. On the other hand, they show marked improvements in $\hat{\theta}$ in every case.

(b) Second refinement: correcting the expected segregation ratio. The corresponding SL estimate of the penetrance f defined in the Models and Methods section was derived by dividing the observed segregation ratio Ψ by an *expected* segregation ratio. The expected segregation ratio used was 0.25 for recessive models, or 0.5 for dominant models. But this denominator is correct only for rare disease alleles. Another refinement of the procedure for deriving corresponding SL analysis models is to use a more accurate expected segregation ratio. We derived formulas for doing this, but found that it had very little effect on the resulting parameter estimates. See Appendix B and Table VI for details.

(3) Is there an exact SL model corresponding to a given 2L model? It might seem that, since the SL models we are considering are simple ones, with a multiplicative structure (as in the tables in Appendix A) and with full penetrance, that it should be possible to find a SL model which would function in exactly the same way in a linkage analysis as the 2L model does. We have already mentioned that one can use reduced penetrance in the SL model to reflect the action of the gene at the unlinked locus in the 2L model. Should it not be possible to find *exactly* the right penetrance? The general answer is no.

In the "2L-based" approach, we used the segregation ratio at the unlinked locus as the SL penetrance, but we pointed out that this was derived from consideration of non-founders only. The "penetrance" for founders is a function of the population gene frequency at the unlinked locus, and not a segregation ratio. Furthermore, exact calculation of the "penetrance" for non-founders, even for nuclear families, will depend on the phenotypes of the parents, the number of children, and how many children are affected. The problem would become even more complicated for extended complex

TABLE VI. Refined SL Parameter Estimates*

Model	Crude		Refined	
	q	f	q	f
RR(1, 3)	0.03	0.64	0.03	0.64
RR(2)	0.04	0.32	0.04	0.33
RR(4, 6)	0.24	0.90	0.27	0.68
RR(5)	0.26	0.74	0.31	0.53
DD(4, 6)	0.05	0.97	0.06	0.92
DD(5)	0.07	0.70	0.08	0.65
DR(1)	0.01	0.48	0.01	0.47
DR(2)	0.01	0.36	0.01	0.36
DR(3)	0.01	0.48	0.01	0.48
RD(1)	0.10	0.97	0.12	0.86

*See Appendix B, Crude estimates repeated from Table II.

pedigrees. There is no direct correspondence between the parameter f of the SL model, and any parameter of the 2L model. There is therefore no analytic procedure for deriving a corresponding SL model that is not just an *approximation* to a given 2L model.

Parenthetically, we note that there *is* an empirical procedure for finding the optimal SL model: find the values of q and f which maximize the lod score in your dataset [Greenberg, 1989; Elston, 1989]. However, this approach partially obviates the advantage of using a SL approximation in the first place, namely, simplicity.

The results of this study have shown that the absence of an exact SL correspondence to epistatic 2L models is not a liability. Provided that Z_{MAX} is the quantity of interest, even our very simple original approach to deriving a corresponding SL model yields excellent approximations. (If $\hat{\theta}$ is the quantity of interest, then one should consider using the 2L-based approach we took in the previous section.)

II. "Between" Variability vs. "Within" Variability

We were concerned about the possibility that a SL analysis might provide a poor approximation to a 2L lod score for *any given dataset* of 20 families, in spite of the overall agreement when the results for each type of analysis are averaged over 50 such datasets. We therefore examined the variability of the difference (dataset by dataset) between the SL and 2L lod scores.

For a single dataset, the SL approximation may differ from the 2L lod score in accordance with the sampling distribution of the mean difference *between* the two. As we noted in the Results, the standard deviations presented in Table III represent variability *within* each type of analysis, i.e., the variability of Z_{MAX} across datasets under one type of analysis (2L or SL).

We found that the magnitude of these "within" standard deviations overwhelms the "between" variability. For instance, for the RR(6) model, the mean dataset by dataset difference *between* the 2L Z_{MAX} and the SL Z_{MAX} is 0.35 with a standard deviation of 0.48. However, the "within" standard deviations are 2.20 and 2.31 for the 2L and SL lod scores respectively (from Table III).

We conclude that, from a practical point of view, the question of the accuracy of the approximation for a single dataset becomes moot. One can obtain a value of Z_{MAX} for a single dataset of 20 nuclear families which differs greatly from the expected Z_{MAX}

for a given generating model, but this will be equally true of the 2L analysis and its SL approximation.

III. Generalizing These Results to Other Genetic Models

We have examined a broad class of 2L generating models, and we believe our results can be further generalized to models with (1) non-zero recombination, (2) additional reduced penetrance, or (3) additional loci. We also note (4) some restrictions on the generalizability of these results.

(1) Non-zero recombination. We chose $\theta = 0$ in order to obtain the cleanest comparison between the two alternative methods of analysis. Moreover, as linkage maps become increasingly dense, it becomes realistic to anticipate mapping genes to markers within 5 cM much, or even most, of the time. However, Vieland et al. [1992], using similar procedures but different generating models, analyzed data simulated with $\theta = 0.1$. That study also found notable agreement between SL and 2L lod scores. We expect that such agreement would hold as well for the case $\theta = 0.5$, i.e., for exclusion of linkage.

(2) Additional reduced penetrance. If we introduce an additional reduced penetrance parameter g into the generating model, the corresponding SL estimate of f will be reduced by g, while the estimate of q will be unchanged. The performance of the SL approximation under additional reduced penetrance (compared to its 2L counterpart) could therefore be expected to be comparable to its performance in the absence of additional reduced penetrance.

(3) Additional loci. We also expect these results to extend to genetic models in which the disease is governed by *more* than two epistatic (mendelian) loci. As discussed earlier, 2L models should be *less* amenable to approximation by SL linkage analyses than multi-locus models. Therefore, we would predict that a SL approximation to an oligogenic model with more than two loci, or to a major gene model with polygenic modifiers, would be *at least as* good as the SL approximation to 2L models considered in this paper.

Note, however, that the accuracy of the SL model as an approximation to an oligogenic model in no way guarantees that a linkage analysis of a complex trait will succeed in establishing linkage: factors such as multiple loci and additional reduced penetrance may reduce the power of a linkage analysis to the point where linkage is undetectable. Our point is simply that this should be more or less *equally* true of the SL model and the oligogenic model it approximates.

(4) Restrictions on generalizability. We also would like to note some restrictions on the generalizability of these results. We have looked only at nuclear families, and linkage analysis of pedigrees might show greater discrepancies in the performance of the SL and 2L analysis models. We have also not considered the performance of the SL approximations under multipoint analysis. Multipoint analysis is, in general, less robust to misspecification of the generating model than is two-point analysis [Risch and Giuffra, 1990], suggesting that SL approximations to 2L generating models may be less appropriate for multipoint analysis. Finally, we have not considered additive or other quantitative genetic effects.

SUMMARY OF CONCLUSIONS

When a disease is or may be caused by two or more mendelian loci in epistasis, a simple SL linkage analysis with a random reduced penetrance parameter should yield

excellent approximations to the true (2L) lod scores. For informative generating models, one may expect SL underestimation of the 2L lod score by 2–15%. But one should bear in mind that $\hat{\theta}$ may not be accurate when a SL analysis is performed for a 2L disease, perhaps particularly under tight linkage, and depending upon the choice of model parameters for the SL analysis. It is also important to correctly specify the mode of inheritance at the linked locus. Where this is an issue, it may be useful to consider two SL models (one recessive, one dominant).

ACKNOWLEDGMENTS

Supported by NIMH grants K21-MH-00884, 5-P50 MH43878, 1-R37 MH 28274; NIH grants DK-31775, NS-27941, NS-21908, DK-31813; and the Klingenstein Foundation.

REFERENCES

Cavalli-Sforza LL, Bodmer WF (1971): "The Genetics of Human Populations." San Francisco: W.H. Freeman, pp. 310–313.

Defrise-Gussenhoven E (1962): Hypotheses de dimerie et de non-penetrance. Acta Genet Stat Med (Basel) 12:65–69.

Elston RC (1989): Man bites dog? The validity of maximizing lod scores to determine mode of inheritance. Am J Med Gen 34:487–488.

Greenberg DA (1981): A simple method for testing two-locus models of inheritance. Am J Hum Genet 33:519–530.

Greenberg DA (1984): Simulation studies of segregation analysis: Application to two-locus models. Am J Hum Genet 36:167–176.

Greenberg DA, Delgado-Escueta AV, Maldonado H, Widelitz H (1988): Segregation analysis of juvenile myoclonic epilepsy. Gen Epi 5:81–94.

Greenberg DA (1989): Inferring mode of inheritance by comparison of lod scores. Am J Med Genet 34:480–486.

Greenberg DA, Hodge SE (1989): Linkage analysis under "random" and "genetic" reduced penetrance. Genet Epi 6:259–264.

Greenberg DA (1990): Linkage analysis assuming a single-locus mode of inheritance for traits determined by two loci: Inferring mode of inheritance and estimating penetrance. Genet Epi 7:467–479.

Lathrop GM, Ott J (1990): Analysis of complex diseases under oligogenic models and intrafamilial heterogeneity by the LINKAGE programs. Am J Hum Genet 47: A188.

Risch N, Giuffra L (1990): Multipoint linkage analysis of genetically complex traits. Am J Hum Genet 47:A197.

Strickberger MW (1976): "Genetics, 2nd Ed." New York: Macmillan, pp. 203–210.

Thomson G (1980): A two locus model for juvenile diabetes. Ann Hum Genet 43:383–398.

Todd JA, Bell JI, McDevitt HO (1987): HLA-DQ_β gene contributes to susceptibility and resistance to insulin-dependent diabetes mellitus. Nature 329:599–604.

Vieland V, Greenberg DA, Hodge SE, Ott J (1992): Linkage analysis of two-locus diseases under single-locus and two-locus analysis models. In MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds): "Issues in Gene Mapping and Detection of Major Genes." Cytogenetics and Cell Genetics: 59:145–146. Basel: S Karger.

Edited by G. P. Vogler

APPENDIX A
2L Model Formulas

2L Penetrance Tables

	RR Models				DR/RD Models				DD Models		
	BB	Bb	bb		BB	Bb	bb		BB	Bb	bb
AA	0	0	0	AA	0	0	1	AA	1	1	0
Aa	0	0	0	Aa	0	0	1	Aa	1	1	0
aa	0	0	1	aa	0	0	0	aa	0	0	0

Definition of the population segregation ratio Ψ:

$$\Psi = \frac{\sum S_i \Psi_i}{\sum S_i \alpha_i}, \quad \text{where } \alpha_i = \begin{array}{l} 1, \text{ when } \Psi_i > 0 \\ 0, \text{ when } \Psi_i = 0 \end{array}$$

$$= \frac{\text{prevalence of disease}}{\text{prevalence of segregating mating types}}$$

where Ψ_i is the proportion of affected offspring for mating type i, and S_i is the frequency of the mating type [Greenberg, 1981].

Formulas for Computing 2L Parameters

Model	K	Ψ
RR	$q_1^2 q_2^2$	$\frac{1}{(1+p_1)^2(1+p_2)^2}$
RD	$q_1^2(1-q_2^2)$	$\frac{1}{(1+p_1)^2(1+q_2^2)}$
DR	$(1-q_1^2)q_2^2$	$\frac{1}{(1+q_1^2)(1+p_2)^2}$
DD	$(1-q_1^2)(1-q_2^2)$	$\frac{1}{(1+q_1^2)(1+q_2^2)}$

where p_1 = frequency of A, q_1 = frequency of a = $1-p_1$
p_2 = frequency of B, q_2 = frequency of b = $1-p_2$
and A, B are the dominant alleles
a, b are the recessive alleles, as in the tables above.

Note: we assume loci are autosomal and in Hardy-Weinberg equilibrium.

APPENDIX B
Second Refinement of the SL Approximation to a 2L Model

In this appendix, we derive formulas adjusting the estimate of q, and therefore f, to allow for segregation ratios other than 0.25 for a recessive disease, or 0.5 for a dominant disease. We then compare the refined parameter estimates to those obtained for the original corresponding SL models.

1. Notation. We start with the same two observed quantities as defined in the text and in Appendix A:

Ψ = population segregation ratio
K = population prevalence.

Then for the "refined" SL model, define

q = frequency of recessive allele
p = 1 − q = frequency of dominant allele
ϕ = SL segregation ratio
f = SL penetrance

2. "Refined" approximation for the SL recessive model. The observations Ψ and K can be expressed as functions of the SL parameters q and f as follows:

$\Psi = f\phi$
$K = q^2 f$

where

$\phi = 1/(1+p)^2$.

Solving the three equations for q and f yields:

$q = 1 - \sqrt{1 - A}$, where $A = \sqrt{K/\Psi}$
$f = K/q_2$.

3."Refined" approximation for the SL dominant model. The observations Ψ and K can be expressed as functions of the SL parameters q and f as follows:

$\Psi = f\phi$
$K = (1-q^2)f$

where

$\phi = 1/(1+q^2)$.

Solving these three equations for q and f yields:

$q = \sqrt{B}$, where $B = \sqrt{1 - (K/\Psi)}$
$f = K/(1-q^2)$.

4. Results. We derived new estimates of q and f for selected models. Table VI compares these "refined" estimates to the original ones.

The refined estimates of q differ hardly at all from the crude ones, while the refined estimates of f differ substantially from the crude estimates in only two cases (for the RR[4] and RR[6] models, which are identical, and the RR[5] model). Thus, the original assumption of $\Psi = 0.25$ (or $\Psi = 0.5$) had relatively little effect on the SL parameter estimates.

In conclusion, although this "refined" approach is more aesthetically pleasing than the original one we used, it makes so little difference in the linkage analyses that we do not recommend its use.