*Am J Hum Genet* 33:396–406, 1981

# Some Epistatic Two-Locus Models of Disease. II. The Confounding of Linkage and Association

SUSAN E. HODGE[1] AND M. ANNE SPENCE

SUMMARY

This paper continues to examine the model discussed in the preceding paper. Specifically, it will be shown how a linkage analysis performed in the presence of a disease-marker association can give rise to erroneous and misleading results.

## INTRODUCTION

The preceding paper [1] has illustrated how a disease-marker association does not necessarily imply the existence of a distinct disease-susceptibility locus tightly linked to the marker locus. An alternative, two-locus model was proposed. Under this hypothesis, the marker locus is causally involved in the disease and interacts epistatically with an additional *unlinked* locus. This model explains the observed association. It also leads to distortions in marker concordance among pairs of affected siblings similar to many reported in the literature for diseases associated with alleles of the HLA complex.

This paper will address directly the question of linkage analyses performed in the presence of associations between the disease and the marker. It will be shown how factors giving rise to the *association* can lead to erroneous *linkage* results.

## MODEL AND METHODS

The model is the same as in [1]. Briefly, expression of the disease is mediated by two loci. One, the "trait" locus, is *not* linked to the marker locus. The other locus is the marker locus itself, where one allele increases susceptibility to the disease. When the marker allele must be

Received April 16, 1980; revised August 26, 1980.

This study was supported by grants MH-33617 from the National Institute of Mental Health and HD-05615, HD-04612, and HD-07032 from the U.S. Public Health Service, and by Mr. and Mrs. Arthur Golden in memory of Margaret Helen Weston.

[1] Both authors: Mental Retardation Research Center, University of California at Los Angeles School of Medicine, 760 Westwood Plaza, Los Angeles, CA 90024.

present in double dose to increase disease susceptibility, the model is called marker-recessive; when only a single dose is required, the model is marker-dominant. When gene action at the marker locus is neither dominant nor recessive, the model is marker-intermediate. The model and its notation are summarized in table 1 in [1].

The effects of this model on a linkage analysis are examined in two examples in this paper. In both examples, data were generated by a two-locus model but then analyzed as if they resulted from a single-locus model, that is, under incorrect assumptions. The lod scores and maximum-likelihood estimates (MLEs) of the recombination fraction $\theta$ were determined by the computer program LIPED [2], using these incorrect assumptions. Thus, the erroneous results to be demonstrated here are the result of erroneous assumptions, not of errors in the lod score method or the program LIPED.

*Example 1. Affected Sib Pairs*

Each sample consisted of a specified number of families, $N$, with both parents unaffected, and exactly two children, both affected, because much of the data in the literature are of this type. All families were analyzed by LIPED *as if* the disease were recessive with complete penetrance. Under these conditions, it is straightforward to compute the exact expected values (averages) of the lod scores and of the MLE of $\theta$. Details of the calculations are now given.

Only three sets of lod scores need to be found: for when the two affected siblings share 2, 1, or 0 genes identical by descent (IBD) at the marker locus (i.e., when they are concordant for both, one or neither marker alleles or haplotypes). These lod scores can be found by LIPED or by hand calculation and are shown in table 1. There are exactly $(N + 2)(N + 1)/2$ ways that $N$ families can be distributed among these three classes [3]. Let the $l$th class consist of those families in which the two affected siblings share $l$ genes IBD. Let $P(l)$ denote the probability of a family being in the $l$th class, and let $x_l$ denote the number of families in the $l$th class. Then the probability of observing any event $E$ consisting of given values of the $x_l$ is

$$\Pr(E) = \Pr(x_2, x_1, x_0) = \frac{N!}{x_2!\, x_1!\, x_0!} P(2)^{x_2} P(1)^{x_1} P(0)^{x_0} ,$$

with the $P(l)$ as given in tables 4, 5, and 6 of [1] for a variety of two-locus models. Moreover, the lod scores observed in this event $E$ are

$$\text{lod}_E(\theta) = \sum_{l=0}^{2} x_l \cdot \text{lod}_l(\theta) ,$$

TABLE 1

LOD SCORES $\text{LOD}_l(\theta)$ FOR A COMPLETELY PENETRANT RECESSIVE DISEASE WHEN BOTH PARENTS ARE UNAFFECTED AND THE FAMILY CONTAINS EXACTLY TWO CHILDREN, BOTH AFFECTED

| NO. GENES SHARED IDENTICAL BY DESCENT AT MARKER LOCUS $l$ | $\theta$ | | | | | |
|---|---|---|---|---|---|---|
| | .01 | .05 | .10 | .20 | .30 | .40 |
| 2 | 0.585 | 0.515 | 0.430 | 0.267 | 0.129 | 0.034 |
| 1 | −1.110 | −0.464 | −0.229 | −0.060 | −0.011 | −0.001 |
| 0 | −2.805 | −1.442 | −0.887 | −0.388 | −0.151 | −0.035 |

where the $\text{lod}_l(\theta)$ are given in table 1 for $l$ genes shared IBD. Thus the exact expected value of the lod score for any value of $\theta$ is found by summing over all possible events:

$$\text{Exp}\,[\text{lod}\,(\theta)] = \sum_{x_2=0}^{N} \sum_{x_1=0}^{N-x_2} \Pr(E) \cdot \text{lod}_E(\theta) \ .$$

("Exp" denotes "expected value.")

Similarly, an MLE of $\theta$ can be found for each possible event. The MLE is taken as one of the six discrete values of $\theta$ being examined. Denote this MLE by $\hat{\theta}_E$; then the exact expected value of $\hat{\theta}$ is

$$\text{Exp}\,(\hat{\theta}) = \sum_{x_2=0}^{N} \sum_{x_1=0}^{N-x_2} \Pr(E) \cdot \hat{\theta}_E \ . \qquad (1)$$

The variance of the MLE is $\text{var}(\hat{\theta}) = \text{Exp}\,(\hat{\theta}^2) - [\text{Exp}\,(\hat{\theta})]^2$, where $\text{Exp}\,(\hat{\theta}^2)$ is found by substituting $\hat{\theta}^2$ for $\hat{\theta}$ in equation (1).

*Example 2. Sibships of Size 3*

To expand sampling possibilities, nuclear families with three children, of whom at least two are affected, were considered. Call these "(3,2) sibships" or families. With three children, the IBD possibilities become immensely more numerous and complex than with two, as well as less illuminating. Hence, another approach was taken. First, only models in which both the trait and marker act recessively, that is, in which $\underset{\sim}{a} = (0,0,1)$ and $\nu_1 > \nu_2 = \nu_3$, were considered (see table 2 in [1]). This choice was made for ease of calculation and because the earlier results suggested a greater effect in marker-recessive than in marker-dominant models. Second, rather than consider all possible (3,2) families, of which there are several hundred, well-defined criteria were used to select the small number of different family structures representing the great majority (over 91%) of such families. Third, rather than the exact average lod scores and MLE for samples consisting of a given number of families, the averages were computed only for a single family, then multiplied by the number of families. (This procedure will be justified in the DISCUSSION.) Details of the calculations are now indicated.

The criteria for selecting a small number of representative family structures fell into two successive steps:

*(1) Parental mating types (m.t.).* The probabilities of 21 possible mating types were found, given that at least two out of three children were affected. These conditional probabilities were found from Bayes' rule:

$$\Pr(\text{m.t.}\,|\,(3,2)\ \text{sibship}) = \frac{\Pr[(3,2)\ \text{sibship}\,|\,\text{m.t.}]\ \Pr(\text{m.t.})}{\sum_{\text{all m.t.}} \Pr[(3,2)\ \text{sibship}\,|\,\text{m.t.}]\ \Pr(\text{m.t.})} \ . \qquad (2)$$

Only those mating types contributing at least 2.5% of the (3,2) families were considered. Using this cutoff reduced the number of mating types to five, while still accounting for at least 93% of all (3,2) sibships.

*(2) Sibship structures.* The "structure" specifies the disease phenotype (affected or normal) and marker type (*MM*, *Mm*, or *mm*) of each of the three children. Of the five mating types chosen above, three were informative for linkage. For these three, the probabilities of all possible sibship structures, given the mating

type and given that only (3,2) sibships were being considered, were found. These conditional probabilities are:

$$\Pr(\text{structure}|\text{m.t., (3,2) sibship}) = \frac{\Pr(\text{structure}|\text{m.t.})}{\Pr[\text{(3,2) sibship}|\text{m.t.}]}, \qquad (3)$$

assuming that only structures of (3,2) type are considered. Again, a cutoff value of 2.5% was used.

This procedure yielded 18 structures informative for linkage, plus a number of structures not informative for linkage, together accounting for at least 91% of all (3,2) sibships. For each of these 18 structures, multiplying equations (2) and (3) yields

$$\Pr(\text{structure}|\text{(3,2) sibship}), \qquad (4)$$

that is, the probability of that family, given only (3,2) families are being considered.

The average lod score for a randomly selected (3,2) family is found as follows: each of the 18 structures found earlier is analyzed by LIPED, under a variety of single-locus models. The resultant lod scores are then weighted by the probabilities in expression (4) and summed over the 18 structures.

To reiterate, in example 1, families of two children each, both affected, are considered. The families are analyzed under one model only: as if the disease were recessive with complete penetrance. The exact expected values of $\hat{\theta}$ and of the lod scores are found as a function of the number of families in a sample. In example 2, each family has three children, at least two of whom are affected. They are analyzed under a variety of genetic models. The average lod score and $\hat{\theta}$ are found for *one* family, then multiplied by the number of families in the sample.

## RESULTS

### *Affected Sib Pairs*

Table 2 shows the average lod scores and the average maximum likelihood estimate $\hat{\theta}$ with its standard error, for samples of $N = 20$ families, generated under

TABLE 2

MEAN LOD SCORES IN AFFECTED SIB PAIRS, AND MEAN AND SE OF $\hat{\theta}$, FOR A SAMPLE OF $N = 20$ FAMILIES

| MODEL | $\theta$ .01 | .05 | .10 | .20 | .30 | .40 | $\hat{\theta} \pm$ SE |
|---|---|---|---|---|---|---|---|
| Marker-dominant $v_1$:$v_3$ | | | | | | | |
| 10:1 | −14.33 | −4.72 | −1.52 | 0.31 | 0.43* | 0.15 | .291 ± .097 |
| 25:1 | −12.03 | −3.41 | −0.63 | 0.75* | 0.61 | 0.19 | .251 ± .085 |
| 1:0 | −10.45 | −2.49 | −0.01 | 1.06* | 0.75 | 0.23 | .227 ± .075 |
| Marker-recessive $v_1$:$v_3$ | | | | | | | |
| 10:1 | −17.39 | −6.50 | −2.72 | −0.28 | 0.17* | 0.08 | .348 ± .109 |
| 25:1 | −7.20 | −0.61 | 1.25 | 1.69* | 1.01 | 0.29 | .185 ± .075 |
| 50:1 | −0.41 | 3.32 | 3.89* | 3.00 | 1.58 | 0.43 | .105 ± .052 |
| 100:1 | 3.00 | 5.28* | 5.21 | 3.66 | 1.86 | 0.50 | .072 ± .036 |
| 1:0 | 5.53 | 6.75* | 6.20 | 4.15 | 2.07 | 0.56 | .053 ± .024 |

NOTE: $r = .10$. Details of model and analysis are in the text. Marker penetrances $v_i$ are defined in table 1 in [1].
* Maximum lod score.

a variety of two-locus models and analyzed as if the disease were recessive with complete penetrance. Three marker-dominant and five marker-recessive models were examined, all with marker gene frequency $r = .10$. Those models that had led to moderate distortions in the sharing of marker types between affected siblings [1] gave rise to suggestive evidence in favor of loose linkage between the marker and a presumed disease susceptibility locus. For example, when the marker penetrance ratio $v_1:v_3$ is 1:0 in a marker-dominant model, the 20 families yield a maximum lod score of 1.06, and the average MLE of $\theta$ is .227. Similarly, a penetrance ratio of 25:1 in a marker-recessive model yields a maximum lod score of 1.69, and the average MLE is .185. The models associated with more striking distortions in marker concordance led to tighter estimates of linkage and to lod scores generally considered tantamount to conclusive proof of linkage: maximum lod scores of 3.89 to 6.75 as the penetrance ratio goes from 50:1 to 1:0 in a recessive model, with the average MLE decreasing from .185 to .053.

To determine whether a value of $N = 20$ accurately reflected what will happen with other sample sizes, the computations were repeated for $N = 30$ and 50 (table 3). Results were also examined for $N = 1$. In all cases, the lod scores were almost exactly proportional to $N$. The MLE of $\theta$ exhibited some bias at lower values of $N$; for $N = 30$ and 50, it converged to a constant value. Note that for $N = 1$, $\hat{\theta}$ was found by interpolating a parabola, as in [4].

*Sibships of Size 3*

Three examples of the recessive-recessive model were considered, with the marker penetrance vector $\underline{v} = (.75, .05, .05)$, $(.75, .03, .03)$, and $(.75, .0075, .0075)$, respectively. Thus $v_1:v_3$ ratios of 15:1, 25:1, and 100:1 were illustrated. In all cases, the trait gene frequency $p = .99$ and the marker gene frequency $r = .1$.

Table 4 shows the five parental mating types comprising at least 2.5% each of the total and gives their frequencies as in equation (2). Note that mating type no. 5, in which both parents are *Aamm*, makes up a considerable portion of (3,2) families when the penetrance ratio $v_1:v_3$ is 15:1 or 25:1 (34% and 20%, respectively). Yet the "associated" marker allele *M* is not present in these families. Note also that none of the five mating types includes an affected parent.

Three of the above mating types are informative for linkage; that is, at least one parent is a double heterozygote. These are nos. 2 (*AaMM* × *AaMm*), 3 (*AaMm* ×

TABLE 3

EFFECTS OF VARYING SAMPLE SIZE *N* ON THE MEAN AND SE OF $\hat{\theta}$ IN TABLE 2

| MODEL | *N* | | | |
|---|---|---|---|---|
| | 1* | 20 | 30 | 50 |
| Marker-dominant 25:1 | .241 | .251 ± .085 | .237 ± .069 | .237 ± .057 |
| Marker-recessive 25:1 | .189 | .185 ± .075 | .179 ± .059 | .179 ± .050 |
| Marker-recessive 50:1 | .117 | .105 ± .052 | .100 ± .044 | .099 ± .031 |

* When $N = 1$, $\hat{\theta}$ is interpolated, as explained in the text.

TABLE 4

PERCENTAGE OF ALL (3,2) SIBSHIPS CONTRIBUTED BY MOST FREQUENT PARENTAL MATING TYPES

| NO. | MATING TYPE | $\underset{\sim}{\nu}$ (.75, .05, .05) | $\underset{\sim}{\nu}$ (.75, .03, .03) | $\underset{\sim}{\nu}$ (.75, .0075, .0075) |
|---|---|---|---|---|
| 1 | *AaMM* × *AaMM* | 1.04% | 1.69% | 2.70% |
| 2 | *AaMM* × *AaMm* | 11.35 | 17.57 | 26.51 |
| 3 | *AaMm* × *AaMm* | 33.34 | 47.32 | 64.08 |
| 4 | *AaMm* × *Aamm* | 15.26 | 8.96 | 0.90 |
| 5 | *Aamm* × *Aamm* | 34.35 | 20.17 | 2.02 |
| | TOTAL | 95.34 | 95.71 | 96.21 |

*AaMm*), and 4 (*AaMm* × *Aamm*). Figure 1 and table 5 show the structures and frequencies, respectively, of the most common sibships resulting from these three mating types. See MODEL AND METHODS for the definition of "most common sibships." Only the conditional frequency of each sibship, given the parental mating type, is shown in the table (equation 3). For the overall frequency, as in expression (4), multiply this conditional frequency by the corresponding mating type frequency from table 4. Note that in only one of these families (no. 2.5) are all three children affected.

Table 6 shows the average lod scores per family, multiplied by 20 so as to facilitate interpretation for a sample of 20 families. The interpolated MLE $\hat{\theta}$ is also given. The linkage analyses were performed under six different models: dominant and recessive, with the nonzero penetrance assuming values of .75, .50, and .05 for each case. In other words, the families were generated by the two-locus recessive-recessive model with $\underset{\sim}{\nu}$ as shown but were analyzed as if the disease were caused by a single locus with the penetrances indicated.

As in the affected sib-pair example, the lod scores increase from suggestive (around 2.0) to conclusive (3.0) as the marker penetrance ratio $\nu_1:\nu_3$ goes from 25:1 to 100:1. The recessive analyses give higher lod scores than the dominant ones. Within a given type of analysis, lowering the penetrance increases the maximum lod score and lowers the MLE of $\theta$.

## DISCUSSION

We have examined a model in which a disease-marker association exists due to an epistatic interaction between the marker locus and an unlinked disease locus. We have illustrated how a linkage analysis performed under these circumstances can give rise to lod scores ranging from suggestive (around 2.0 or higher) to conclusive (3.0) in modest-sized samples of 20 small nuclear families. Moreover, the resultant estimate of the recombination fraction $\theta$ may be fairly large (25%–30%) or as low as 0.

From table 6, the lod scores are higher when analyzed under a recessive model than under a dominant one. However, both models give positive results, which are not highly dependent on the assumed penetrance. Thus, positive results under a variety of assumed single-locus models do not guarantee that any of these models is

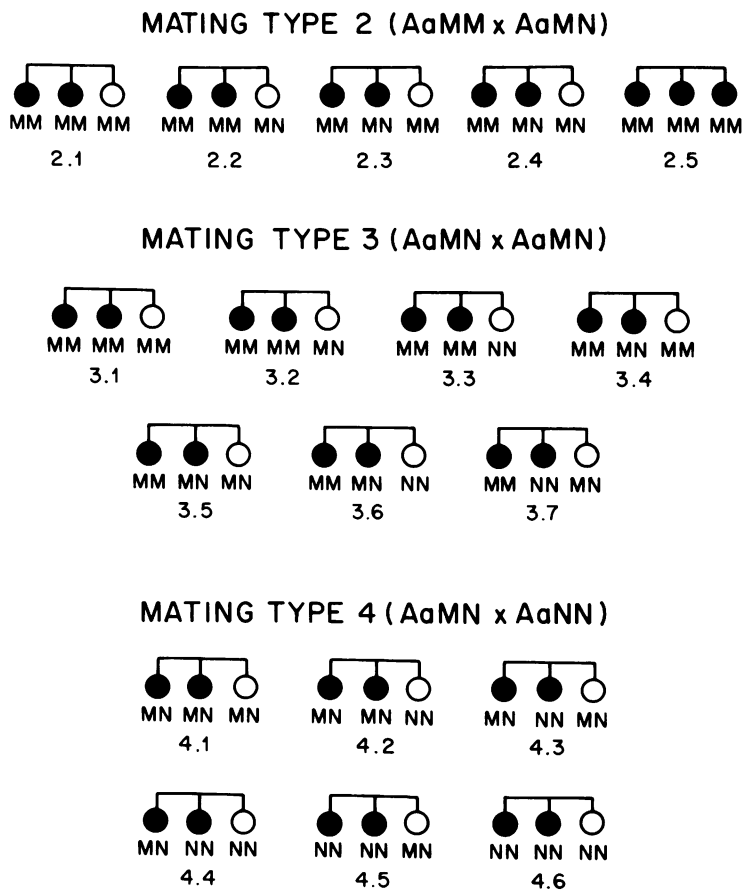


FIG. 1.—Phenotypic appearance of the 18 most common (3,2) structures, broken down by mating type of parents.

correct or that a distinct disease-susceptibility locus is actually linked to the marker locus. This point has already been made in [5].

The misleading findings of linkage are presumably caused by two factors. First, the observed $\hat{\theta}$ represents a kind of "average" of the 50% recombination fraction between the trait and marker loci and the "zero" recombination fraction between the marker locus and itself. Second, the association itself clearly inflates linkage results: affected persons tend to share marker types at the *population* level, because of the association; this increased sharing appears on the *family* level, where it simulates linkage.

Consider the alternative "hitchhiking" model in which the disease is caused by a single locus fairly close to the marker (but with $\theta > 0$) and simultaneously in linkage disequilibrium with it. It is reasonable to speculate that even in this case a linkage analysis would still tend to inflate lod scores and bias the estimate of $\theta$

TABLE 5

CONDITIONAL FREQUENCIES, GIVEN PARENTAL MATING TYPE, AND OVERALL FREQUENCIES AMONG ALL (3,2) FAMILIES OF THE 18 MOST FREQUENT (3,2) SIBSHIPS

| MATING TYPE | SIBSHIP NO. | $\underset{\sim}{\nu}$ (.75, .05, .05) | (.75, .03, .03) | (.75, .0075, .0075) |
|---|---|---|---|---|
| 2 | 2.1 | 38.26% | 40.17% | 42.51% |
| | 2.2 | 46.50 | 49.07 | 52.22 |
| | 2.3 | 5.10 | 3.21 | * |
| | 2.4 | 6.20 | 3.93 | * |
| | 2.5 | 2.94 | 3.09 | 3.27 |
| | Total | 99.00 | 99.47 | 98.00 |
| 3 | 3.1 | 14.66 | 16.78 | 19.78 |
| | 3.2 | 35.62 | 41.00 | 48.61 |
| | 3.3 | 17.81 | 20.50 | 24.30 |
| | 3.4 | 3.91 | 2.68 | * |
| | 3.5 | 9.50 | 6.56 | * |
| | 3.6 | 4.75 | 3.28 | * |
| | 3.7 | 4.75 | 3.28 | * |
| | Total | 91.00 | 94.08 | 92.69 |
| 4 | 4.1 | 12.45 | 12.47 | 12.49 |
| | 4.2 | 12.45 | 12.47 | 12.49 |
| | 4.3 | 24.90 | 24.94 | 24.98 |
| | 4.4 | 24.90 | 24.94 | 24.98 |
| | 4.5 | 12.45 | 12.47 | 12.49 |
| | 4.6 | 12.45 | 12.47 | 12.49 |
| | Total | 99.60 | 99.76 | 99.92 |

* Less than 2.5%.

downward also, because of the effects of the association. For example, if the true $\theta$ were 3%–5%, the estimated $\theta$ might be 0.

Four decisions relating to design of this study will now be discussed.

(1) The decision to examine nuclear families rather than large pedigrees was based on the following reasoning. Much of the available data on the diseases of interest are in the form of nuclear families. It may be that the effects on linkage analysis would be attenuated in large pedigrees (although this is not certain), but qualitatively they should remain the same. Thus it was felt that the results described here can be generalized to more complex family structures. Similarly, large-scale simulation, although much more expensive, would not have provided additional insight into the problem.

(2) Choice of the recessive-recessive model in example 2 was dictated by several considerations. It was already known that in the case of sib pairs, both the IBD distribution [1] and the lod score analysis (example 1) are distorted most markedly in marker-recessive models. Trait-recessive models are easier to calculate than trait-dominant ones and are certainly appropriate for nuclear families. Moreover, since the characteristics of the trait locus had *no* effect on the linkage analysis of affected sib pairs, it was reasonable to assume they would have little effect on three-child families.

TABLE 6

AVERAGE LOD SCORES PER FAMILY, MULTIPLIED BY 20, AND INTERPOLATED MLE OF RECOMBINATION FRACTION $\theta$

| GENERATING MODEL $\underset{\sim}{\nu}$ | MODEL USED IN LINKAGE ANALYSIS | PENETRANCE | $\theta$ .01 | .05 | .10 | .20 | .30 | .40 | INTERPOLATED $\hat{\theta}$ |
|---|---|---|---|---|---|---|---|---|---|
| .75, .05, .05 (15:1) | Dominant | .75 ..... | 0.76 | 0.88* | 0.84 | 0.62 | 0.32 | 0.10 | .066 |
| | | .50 ..... | 1.00 | 1.04* | 0.94 | 0.64 | 0.32 | 0.08 | .045 |
| | | .05 ..... | 1.06* | 1.06* | 0.96 | 0.64 | 0.32 | 0.08 | .030 |
| | Recessive | .75 ..... | −1.08 | 0.86 | 1.32* | 1.14 | 0.62 | 0.18 | .138 |
| | | .50 ..... | −0.78 | 0.98 | 1.38* | 1.16 | 0.62 | 0.18 | .134 |
| | | .05 ..... | −0.48 | 1.04 | 1.38* | 1.14 | 0.62 | 0.18 | .130 |
| .75, .03, .03 (25:1) | Dominant | .75 ..... | 1.82* | 1.80 | 1.62 | 1.10 | 0.56 | 0.16 | .023 |
| | | .50 ..... | 2.10* | 2.00 | 1.74 | 1.14 | 0.56 | 0.16 | 0 |
| | | .05 ..... | 2.16* | 2.02 | 1.74 | 1.12 | 0.54 | 0.14 | 0 |
| | Recessive | .75 ..... | 1.60 | 2.92 | 2.94* | 2.12 | 1.10 | 0.30 | .078 |
| | | .50 ..... | 1.94 | 3.08* | 3.03 | 2.14 | 1.10 | 0.30 | .077 |
| | | .05 ..... | 2.18 | 3.12* | 3.02 | 2.12 | 1.06 | 0.28 | .071 |
| .75, .0075, .0075 (100:1) | Dominant | .75 ..... | 3.92* | 3.50 | 2.98 | 1.92 | 0.96 | 0.26 | 0 |
| | | .50 ..... | 4.28* | 3.76 | 3.14 | 1.94 | 0.96 | 0.26 | 0 |
| | | .05 ..... | 4.30* | 3.76 | 3.12 | 1.92 | 0.92 | 0.24 | 0 |
| | Recessive | .75 ..... | 8.08* | 7.22 | 6.10 | 3.88 | 1.90 | 0.50 | 0 |
| | | .50 ..... | 8.40* | 7.44 | 6.22 | 3.90 | 1.90 | 0.50 | 0 |
| | | .05 ..... | 8.42* | 7.42 | 6.18 | 3.84 | 1.86 | 0.50 | 0 |

NOTE: $r = .1$ and $p = .99$.
* Maximum lod score.

(3) In example 2, there were two justifications for including only the small number of possible (3,2) structures making up the great majority of the population. First, a priori the deleted structures exhibit no particular or consistent differences from the included ones. Second, the results agree qualitatively with those from example 1, where all possible structures were considered.

(4) The average lod scores for a sample of, say, 20 families does not equal the average lod score for one family multiplied by 20. (The expected value of a function does not equal the function of the expected value.) Nor does the MLE of $\theta$ for a sample of 20 families equal the MLE for one family. However, the discrepancy between the correct and incorrect values was small in the affected sib-pair case (example 1), as discussed in the results and shown in table 3. Hence, it was felt that in the (3,2) sibships it was sufficient to calculate the results for single families and that this procedure would indicate results adequately for the purposes of this paper.

It might be argued that if the associated marker allele is not found in all patients or in all families segregating the disease, then that allele cannot itself be causally implicated in the disease. However, such is not the case in this model, where the associated allele increases susceptibility but is neither necessary nor sufficient for expression of the disease. From table 4, in a recessive-recessive disease with a 15:1 or 25:1 penetrance ratio at the marker locus, one-third to one-fifth of all (3,2) families do not possess the *M* allele at all. Yet persuasive evidence in favor of linkage still appears (tables 2 and 6).

However, note that those families without *M* should, on the average, fail to simulate linkage, whereas under the "hitchhiking" model they would continue to give evidence supporting linkage. Analyzing these families separately, as done in [5] for multiple sclerosis, may represent one approach to resolving the linkage-association problem.

Of current interest are associations between the HLA marker system and a number of diseases. Several studies have examined these diseases for linkage with HLA, assuming a single-locus model for the disorder. Juvenile-type (insulin-dependent) diabetes [6, 7], multiple sclerosis [5], hemochromatosis [8], and hypertrophic cardiomyopathy [9, 10] have been analyzed in this fashion. However, the issue of association was not always addressed in these studies. It is noteworthy, too, that both loose and tight linkage have been found in these analyses.

A particularly striking association is that between ankylosing spondylitis (AS) and the *B27* allele of the HLA system. The association is much stronger than for many other diseases, with at least 90% of AS patients exhibiting the B27 antigen vs. 5%–10% in controls [11]. The same allele is associated with AS in a variety of ethnic groups [12], whereas the "hitchhiking" theory predicts that although the association will be found in all populations, the particular allele will vary from group to group [8]. Recent findings [13, 14] lend support to a "molecular mimicry" explanation for AS, whereby B27 itself may cross-react with the *Klebsiella pneumoniae* antigen to help cause the disease. This disorder might represent a particularly appropriate example to start with in any attempt to resolve the issues raised in this paper.

It might appear that the distinction being drawn here between (1) tight linkage between marker and susceptibility loci and (2) identity of the two loci (pleiotropy)

were merely semantic. It has always been known that formal linkage analysis cannot make this distinction. Only when a crossover has been confirmed can the existence of two separate loci be considered proven. However, confirming crossovers in these diseases in which modes of inheritance are unknown is not possible at present. In any case, we feel the above distinction is not trivial, but is important to our understanding of both marker locus action and disease pathophysiology.

In conclusion, we would like to sound a note of caution when linkage analyses are attempted in the presence of disease-marker associations. We have not demonstrated that the kind of model proposed here is correct or that the "hitchhiking" theory is wrong. Moreover, different diseases will probably have different explanations. We have demonstrated, however, that this alternative two-locus model can exhibit many of the same features as the "hitchhiking" model [1] and can simulate the presence of a linked susceptibility locus where none exists. The next step is to develop methods for making the preliminary biological decision as to whether or not the marker is causally involved in the disease.

## REFERENCES


1. Hodge SE: Some epistatic two-locus models of disease. I. Relative risks and identity-by-descent distributions in affected sib pairs. *Am J Hum Genet* 33:381–395, 1981
2. Ott J: Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 26:588–597, 1974
3. Feller W: *An Introduction to Probability Theory and Its Applications*, vol I. New York, John Wiley, 1968, p 38
4. Morton NE: The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am J Hum Genet* 8:80–96, 1956
5. Tiwari J, Hodge SE, Spence MA, Terasaki PI: HLA and the inheritance of multiple sclerosis: linkage analysis of 72 pedigrees. *Am J Hum Genet* 32:103–111, 1980
6. Suarez B, Hodge SE, Reich T: Is juvenile diabetes determined by a single gene closely linked to HLA? *Diabetes* 28:527–532, 1979
7. Barbosa J, Chern MM, Noreen H, Anderson VE, Yunis EJ: Analysis of linkage between the major histocompatibility system and juvenile, insulin-dependent diabetes in multiplex families. *J Clin Invest* 62:492–495, 1978
8. Kravitz K, Skolnick M, Cannings C, et al.: Genetic linkage between hereditary hemochromatosis and HLA. *Am J Hum Genet* 31:601–619, 1979
9. Darsee JR, Heymsfeld SB, Nutter DO: Hypertrophic cardiomyopathy and human leukocyte antigen linkage. *N Engl J Med* 300:877–881, 1979
10. Hodge SE, Spence MA, Cederbaum SD: Hypertrophic cardiomyopathy. (Letter to the Editor.) *N Engl J Med* 301:442–443, 1979
11. Kidd KK, Bernoco D, Carbonara AO, Daneo V, Steiger U, Cepellini R: Genetic analysis of HLA-associated diseases: the "illness-susceptibility" gene frequency and sex ratio in ankylosing spondylitis, in *HLA and Disease*, edited by Dausset J, Svejgaard A, Baltimore, Williams and Wilkins, 1977, pp 72–80
12. Sachs JA, Brewerton DA: HLA, ankylosing spondylitis and rheumatoid arthritis. *Br Med Bull* 34:275–278, 1978
13. Ebringer RW, Cawdell DR, Cowling P, Ebringer A: Sequential studies in ankylosing spondylitis: association of *Klebsiella pneumoniae* with active disease. *Ann Rheum Dis* 37:146–151, 1978
14. Seager K, Bashir HV, Geczy AF, Edmonds J, de Vere-Tyndall A: Evidence for a specific B27-associated cell surface marker on lymphocytes of patients with ankylosing spondylitis. *Nature* 277:68–70, 1979