Am J Hum Genet 38:599–616, 1986

Detecting Linkage for Genetically Heterogeneous Diseases and Detecting Heterogeneity with Linkage Data

L. L. CAVALLI-SFORZA[1] AND MARY-CLAIRE KING[2]

SUMMARY

Interest in searching for genetic linkage between diseases and marker loci has been greatly increased by the recent introduction of DNA polymorphisms. However, even for the most well-behaved Mendelian disorders, those with clear-cut mode of inheritance, complete penetrance, and no phenocopies, genetic heterogeneity may exist; that is, in the population there may be more than one locus that can determine the disease, and these loci may not be linked. In such cases, two questions arise: (1) What sample size is necessary to detect linkage for a genetically heterogeneous disease? (2) What sample size is necessary to detect heterogeneity given linkage between a disease and a marker locus?

We have answered these questions for the most important types of matings under specified conditions: linkage phase known or unknown, number of alleles involved in the cross at the marker locus, and different numbers of affected and unaffected children. In general, the presence of heterogeneity increases the recombination value at which lod scores peak, by an amount that increases with the degree of heterogeneity. There is a corresponding increase in the number of families necessary to establish linkage.

For the specific case of backcrosses between disease and marker loci with two alleles, linkage can be detected at recombination fractions up to 20% with reasonable numbers of families, even if only half the families carry the disease locus linked to the marker. The task is easier if more than two informative children are available or if phase is known. For recessive diseases, highly polymorphic markers with four different alleles in the parents greatly reduce the number of families required.

Received June 12, 1985; revised September 19, 1985.

This work was supported in part by grants GM-28428, CA-27632, and AM-34942 from the National Institutes of Health.

[1] Stanford University, Stanford, CA 94305.

[2] University of California, Berkeley, CA 94720.

It is possible to detect heterogeneity by comparing the maximum likelihood value obtained as a function of both recombination and heterogeneity with the likelihood value obtained on the assumption of no heterogeneity. The numbers of families necessary to establish heterogeneity are minimal near 50% heterogeneity. For recessive diseases with phase unknown, backcrosses for the disease and marker loci do not allow separate estimation of linkage and heterogeneity and, therefore, preclude testing heterogeneity, unless the number of affected children per family is at least four, an inevitably rare situation. This impasse is overcome if phase is known (a rare event for rare recessive diseases) or if a highly polymorphic marker is available, yielding $A_1A_2 \times A_3A_4$ parental genotypes. For rare dominant diseases, for which most matings are backcrosses involving only one informative parent, heterogeneity can be detected only if phase is known or if at least four children per family are available. Linkage can be tested in the presence of undetected heterogeneity, but if heterogeneity exists, the estimated recombination value for linkage will be too high.

INTRODUCTION

The availability of many genetic markers that are highly polymorphic at the DNA level has revolutionized genetic analysis, so that it is now possible, in principle, to map the genes responsible for any genetic disease. However, most genetically influenced common diseases have complications, including genetic heterogeneity, incomplete penetrance, variable age of onset, and occurrence of nongenetic cases (phenocopies), that make genetic analysis more difficult. The problem of genetic heterogeneity is especially crucial to address whenever linkage between a disease and a marker is detected and used for genetic counseling. The most direct way to prove genetic heterogeneity would be to map all genes that independently determine a given disease. One approach is to sample a few very large pedigrees, each sufficiently informative to show linkage independently. Different linkages in different pedigrees would establish heterogeneity in this case. However, for many diseases, enough large families may not be available to detect in this way heterogeneity of practical importance. To demonstrate heterogeneity directly with smaller families would be possible if all disease genes have closely linked markers, but the mapping techniques are so demanding that this will take a long time. Meanwhile, it is of interest to test statistically for heterogeneity and anticipate the number of families necessary to detect linkage when heterogeneity is present and the information comes from a sample of small families.

In this analysis, we address the problem presented to linkage studies by genetic heterogeneity—the existence of two or more unlinked loci that determine disease susceptibility. The problem was originally addressed by Morton [1] and Smith [2]. Ott reformulated the Smith approach as a likelihood ratio test [3] and compared the Smith and Morton approaches [4]. These methods were used to test for genetic heterogeneity of bipolar-related affective illness [5] and insulin-dependent diabetes [6].

For fully penetrant diseases, without phenocopies, we are interested in two questions: (1) What sample size is necessary to detect linkage for a genetically heterogeneous disease? and (2) What sample size is necessary to detect heterogeneity given linkage between a disease allele and a marker locus? For simplicity, we confine our attention to the most frequent and most informative matings. Let

$A_1, A_2 \ldots$ = alleles at marker locus A;
D = disease locus linked to marker locus A; f_1 is the frequency of the allele responsible for disease at this locus. Disease alleles at other loci, not linked to A or D, have frequencies f_2, $f_3 \ldots$. When necessary to specify, for a recessive disease, affected individuals will be *dd*, and, for a dominant disease, *dd* will be normal.
α = true proportion of cases in the population due to disease locus D. If the disease alleles are all recessive, then approximately $\alpha = (f_1^2)/(\Sigma f_i^2)$. If the disease alleles are all dominant, then again, approximately $\alpha = (f_1)/(\Sigma f_i)$.
a = estimated proportion of cases in the population due to D; depends on r;
p = proportion of affected sib pairs discordant for genotypes at the marker locus A;
θ = true probability of recombination between D and marker locus A;
r = estimated θ from sample of families, depends on a;
N = number of families required to detect linkage or heterogeneity between D and A;
n = number of children to be considered in each family (for a dominant disease, n = all children; for a recessive disease, n = number of affected children);
i = number of children per family with recombinant genotypes for A and D when linkage phase of A and D is known, or one of two parental marker types if phase is unknown;
ϕ = phase for marker loci with two alleles ($\phi = 1$ if phase is known, $\phi = \frac{1}{2}$ if phase is unknown); and
ψ = phase for marker loci with many alleles ($\psi = 1$ if phase is known, and $\psi = 0$ if phase is unknown).

NUMBERS OF FAMILIES NECESSARY TO DETECT LINKAGE

Marker Locus with Two Alleles

The number of families N necessary to detect linkage between D and A at θ is calculated as the lod score associated with linkage (say 3.0) divided by the expected lod score for one family. Consider families that are backcrosses at the marker locus—that is, for a recessive disease, each family has $A_1A_2 \times A_kA_k$ parents ($k = 1, 2, \ldots$); for a dominant disease, the affected parent is A_1A_2

and the unaffected parent A_kA_k. Following the approach of Morton [1], if the linkage phase of the marker locus and the disease locus is known, then for a family with disease allele at locus D, i children with recombinant genotypes and $(n - i)$ children with nonrecombinant genotypes, the lod score for linkage of D and A at recombination fraction r is

$$\log_{10} \frac{(1 - r)^{n-i} r^i}{\left(\frac{1}{2}\right)^n} .$$

If the linkage phase of the marker locus and the disease locus is unknown, then the lod score at r for a family, with disease linked to A, with i children of one phase and $(n - i)$ children of the opposite phase, is

$$\log_{10} \frac{\frac{1}{2}(1 - r)^{n-i} r^i + \frac{1}{2}(1 - r)^i r^{n-i}}{\left(\frac{1}{2}\right)^n} . \tag{1}$$

For a family with disease due to a locus not linked to A, the lod score is

$$\log_{10} \frac{\left(\frac{1}{2}\right)^n}{\left(\frac{1}{2}\right)^n} = 0 .$$

A general formula, for known or unknown phase, for the lod score at any a and r for a family with i children of one phase (recombinant if phase known) at the A and D loci and $(n - i)$ children of the alternate phase (nonrecombinant if phase known) at the A and D loci is

$$\text{lod}(i) = \log_{10} \frac{\phi a(1 - r)^{n-i} r^i + (1 - \phi) a(1 - r)^i r^{n-i} + (1 - a)\left(\frac{1}{2}\right)^n}{\left(\frac{1}{2}\right)^n} .$$

The expected lod score for a family with n children depends on both lod (i) and the expected proportion P(i) of families with i recombinant and $(n - i)$ nonrecombinant children. The distribution of family types depends on the true α and θ for linkage of D to A. For families with disease due to D or d, the proportion of families with i recombinant and $(n - i)$ nonrecombinant children is

$$\binom{n}{i}(1 - \theta)^{n-i}\theta^i .$$

For families with disease at an unlinked locus, the proportion of families with i recombinant and $(n - i)$ nonrecombinant children is

$$\binom{n}{i}\left(\frac{1}{2}\right)^n .$$

For proportion α of families with disease linked to A, and $(1 - \alpha)$ families with unlinked disease loci,

$$P(i) = \binom{n}{i}\left[\alpha(1 - \theta)^{n-i}\theta^i + (1 - \alpha)\left(\frac{1}{2}\right)^n\right] . \quad (2)$$

The expected lod score for a family of n children of $A_1A_2 \times A_kA_k$ parents is therefore

$$E(L) = \sum_{i=0}^{n} P(i)\text{lod}(i) . \quad (3)$$

Figure 1 illustrates how E(L) varies as a function of r and a for $\theta = .1$, $\alpha = .5$, with $n = 3$ and phase unknown, and with $n = 4$ and phase known. As we shall also discuss further, with $n = 2$ or 3 and phase unknown, the peak lod score is independent of a (fig. 1), and therefore one cannot estimate α separately from θ using maximum likelihood.

The expected lod score for N families is equal to N E(L) with a maximum value of $N L_{max}$ over all θ and α. Therefore, to obtain a total lod score of three, the number of families required, on average, is $N = 3/L_{max}$. Table 1 indicates the number of families with $n = 2$, $n = 3$, and $n = 4$ children necessary to obtain a lod score of 3.0 for various recombination fractions, in the presence of genetic heterogeneity for phase unknown. As table 1 indicates, N is larger for

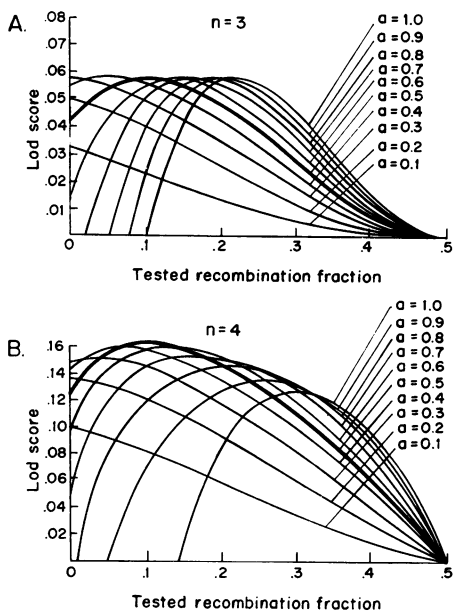

FIG. 1.—Expected lod score as a function of recombination fraction at various levels of genetic heterogeneity for matings $DdA_1A_2 \times ddA_kA_k$ (dominant disease) or $DdA_1A_2 \times DdA_kA_k$ (recessive disease). True $\alpha = .50$, $\theta = .10$. *A*, Phase unknown, three informative children per family (all children for a dominant disease; affected children for a recessive disease). *B*, Phase known, four informative children per family.

TABLE 1

NO. FAMILIES REQUIRED TO DETECT LINKAGE AT 3.0, GIVEN $A_1A_2 \times A_kA_k$ PARENTS, PHASE UNKNOWN

α	θ .001	.05	.10	.15	.20
			$n = 2$		
1.0	10	18	31	55	104
.9	14	23	39	69	129
.8	19	30	50	88	164
.7	26	40	66	115	215
.6	36	56	91	157	294
.5	53	82	133	228	424
.4	85	129	208	357	664
.3	152	232	372	637	1,182
.2	346	524	841	1,436	2,662
.1	1,390	2,103	3,370	5,752	>10,000
			$n = 3$		
1.0	5.1	8.4	14	23	41
.9	6.7	10	17	28	51
.8	8.7	13	21	35	63
.7	11	17	27	45	82
.6	16	23	37	61	110
.5	22	33	52	86	156
.4	34	51	80	133	240
.3	59	88	138	231	420
.2	129	192	302	508	928
.1	492	737	1,169	1,978	3,635
			$n = 4$		
1.0	3.4	5.3	8.3	13	24
.9	4.3	6.6	10	16	29
.8	5.5	8.2	13	20	36
.7	7.0	11	16	26	46
.6	9.3	14	21	35	61
.5	13	19	30	48	86
.4	19	28	44	73	130
.3	32	48	75	124	223
.2	65	99	158	265	481
.1	234	361	583	996	1,837

lower values of α. These figures are appropriate for estimating necessary sample sizes for diseases like cystic fibrosis or neurofibromatosis, given various levels of genetic heterogeneity and unknown phase.

Family sizes are based on all children for dominant diseases, but only affected children for recessive diseases, because unaffected children add little to the lod score for a recessive disease. Suppose parents carrying the recessive disease allele are $A_1A_2 \times A_kA_k$ and that the disease allele is linked to A_2 at distance θ. Then, among unaffected children, the proportion A_1 will be $(2 - \theta)/$

3 and the proportion A_2 will be $(1 + \theta)/3$. The probability that the number of A_1 unaffected children in a family is j and the number of A_2 unaffected children is $(m - j)$ is

$$P(j) = \binom{m}{j}\left(\frac{2 - \theta}{3}\right)^{m-j}\left(\frac{1 + \theta}{3}\right)^{j} . \tag{4}$$

The lod score contributed by the unaffected children in such a mating of carriers is

$$\text{lod}(j) = \log_{10} \frac{\phi a\left(\frac{2 - r}{3}\right)^{m-j}\left(\frac{1 + r}{3}\right)^{j} + (1 - \phi)a\left(\frac{1 + r}{3}\right)^{m-j}\left(\frac{2 - r}{3}\right)^{j} + (1 - a)\left(\frac{1}{2}\right)^{m}}{\left(\frac{1}{2}\right)^{m}} \tag{5}$$

and

$$E(L_u) = \sum_{j=0}^{m} P(j)\text{lod}(j) . \tag{6}$$

Figure 2 is a numerical example showing that the lod score increase is small when unaffected children are added. For example, adding four unaffected children to a family in which two parents and two affected children were tested increases the expected lod score by only about 13%, despite doubling the number of subjects. When a recessive disease is common enough that it is possible to choose only families with at least two living affected children, it is more efficient to limit analysis to these multiply-affected families, regardless of the number of unaffected children in each. It does not generally appear worthwhile to search specifically for families with many unaffected children.

If the linkage phase of D and A is known, then fewer families are required to detect linkage. In practice, this usually means the disease will be a dominant, with grandparents available. Because in this case affected and unaffected children are equally informative, families of $n = 4$ or more children may be available and useful. Table 2 indicates required sample sizes for families with an A_1A_2 affected parent and an A_kA_k unaffected parent, known phase, and two, three, or four children per family. Clearly, larger families are more informative, and detecting linkage requires more families for smaller α. Knowing linkage phase, however, considerably increases the amount of information per family. For example, at $\theta = .1$ and $\alpha = .7$, detecting linkage requires 20 families of $n = 2$ children if phase is known and 66 families if phase is unknown.

A Particular Problem of Recessive Disease, n = *2 or 3*

For a recessive disease like cystic fibrosis, linkage phase is usually unknown. Families with one affected child have very little information for linkage, but families with two, and occasionally three, affected children may be available. We have already seen that additional unaffected children add proportionately

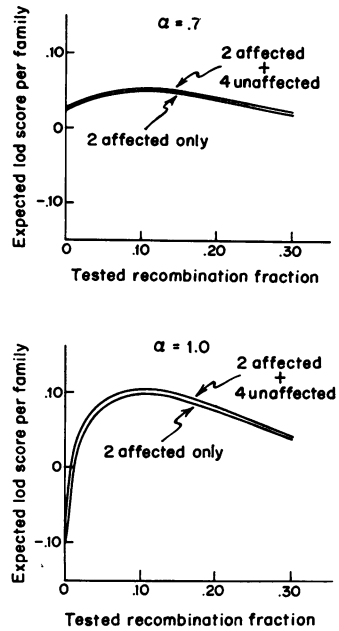

FIG. 2.—Effect on lod score of including unaffected children in linkage analysis of recessive disease, given two affected children per family, true $\theta = .10$, and phase unknown. Expected lod scores per family if no unaffected children are tested (*lower curve of each pair*) and if four unaffected children are also tested in each family (*upper curve of each pair*). Comparisons are given for true $\alpha = .7$ and 1.0, when lod scores are calculated at the true heterogeneity using equations (3) and (4).

much less information. Therefore, the most common informative families of type $A_1A_2Dd \times A_kA_kDd$ for recessive diseases have two affected children. In such families, phase is almost always unknown. These families can be distinguished into two types: concordant sib pairs (i.e., both affected children are A_1 or both are A_2) and discordant sib pairs.

The proportion of discordant sib pairs, p, can be used for testing linkage. The probability that a pair of affected children will be concordant at the marker locus A is $(1 - \theta)^2 + \theta^2$, and the probability the affected pair will be discordant A is $2\theta(1 - \theta)$. For families with disease caused by other loci, the probability that the affected pair of children will be concordant at A is ½ and the probability of discordance is ½.

Therefore:

$$p_{\theta,\alpha} = 2\alpha\theta(1 - \theta) + \frac{1 - \alpha}{2} \tag{7}$$

for families with two affected children. If the disease and marker loci are unlinked, then $\theta = ½$, so $p = ½$ for all α; for complete linkage, $p = 0$.

A given value of p corresponds to an infinite family of θ, α values, and therefore estimation of linkage from p cannot separate α from θ. With only

TABLE 2

NO. FAMILIES REQUIRED TO DETECT LINKAGE AT 3.0, GIVEN $A_1A_2 \times A_kA_k$ PARENTS AND PHASE KNOWN

α	θ				
	.001	.05	.10	.15	.20
			$n = 2$		
1.0	5.0	7.0	9.4	13	18
.9	6.7	8.9	12	16	22
.8	8.7	11	15	20	28
.7	12	15	20	26	37
.6	16	20	26	36	50
.5	22	28	37	51	71
.4	34	44	58	78	110
.3	59	76	101	137	193
.2	129	166	221	302	427
.1	493	638	854	1,173	1,672
			$n = 3$		
1.0	3.4	4.7	6.3	8.5	12
.9	4.3	5.8	7.8	11	15
.8	5.5	7.3	10	13	19
.7	7.0	9.4	13	17	24
.6	9.3	13	17	23	32
.5	13	17	23	32	45
.4	19	25	34	48	69
.3	31	42	58	82	119
.2	65	88	123	175	257
.1	233	321	453	654	975
			$n = 4$		
1.0	2.6	3.5	4.7	6.4	10
.9	3.2	4.4	5.9	7.9	11
.8	3.9	5.4	7.2	10	14
.7	4.9	6.8	9.2	13	18
.6	6.4	8.8	12	17	24
.5	8.6	12	16	23	33
.4	12	17	24	33	49
.3	20	28	39	56	83
.2	38	55	79	116	175
.1	128	186	275	415	643

families with two affected children, $A_1A_2 \times A_kA_k$ parents, and phase unknown, it is not possible to estimate θ and α separately. Figure 3*A* shows the curves of θ, α pairs corresponding to linkage for a set of chosen p values for $n = 2$.

For recessive disease families with three affected children of $A_1A_2 \times A_kA_k$ parents carrying a disease allele linked to *A*, with phase unknown, the probability of concordance at *A* of all three children is $(1 - \theta)^3 + \theta^3$, and the probability of discordance, $1 - [(1 - \theta)^3 + \theta^3]$. For families with three children affected

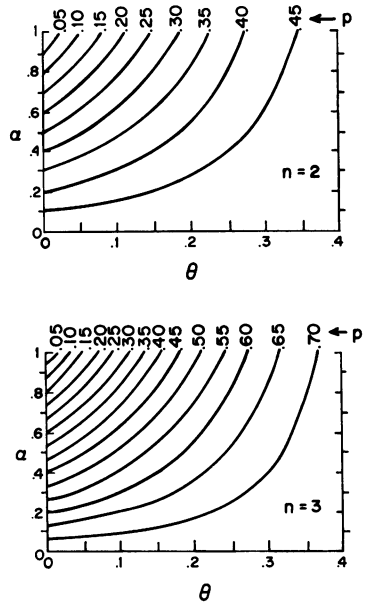

FIG. 3.—Range of values of heterogeneity and recombination fraction consistent with linkage, given observed proportion p of affected sib-pairs discordant at the marker locus, for $DdA_1A_2 \times DdA_kA_k$ matings. *Top*, $n = 2$, $\alpha = (\frac{1}{2} - p)/[\frac{1}{2} - 2\theta(1 - \theta)]$. *Bottom*, $n = 3$, $\alpha = (\frac{3}{4} - p)/[\frac{3}{4} - 3\theta(1 - \theta)]$.

with disease due to a locus unlinked to A, the probability of concordance is $\frac{1}{4}$ and the probability of discordance is $\frac{3}{4}$. Therefore,

$$p_{\theta,\alpha} = \alpha[1 - (1 - \theta)^3 - \theta^3] + (1 - \alpha)\left(\frac{3}{4}\right) . \quad (8)$$

If the disease and marker loci are unlinked, then $\theta = \frac{1}{2}$ and $p = \frac{3}{4}$. Figure 3*B* indicates pairs of θ,α values corresponding to observed proportions of discordant sibs for $n = 3$. As before, separate estimation of α and θ is impossible.

The number of families necessary to detect linkage using discordant sibs can be estimated from equations (7) and (8). The appropriate χ^2 test (1 degree of freedom) for linkage based on observed $p_{\theta,\alpha}$ is

$$\chi^2 = \frac{(Np_{\theta,\alpha} - Np_{\theta=.5})^2}{Np_{\theta=.5}} + \frac{[N(1 - p_{\theta,\alpha}) - N(1 - p_{\theta=.5})]^2}{N(1 - p_{\theta=.5})} .$$

The number of families with two affected children necessary to reach a given significant $\bar{\chi}^2$ is $N_2 - \bar{\chi}^2/16\alpha^2[\theta(1 - \theta) - \frac{1}{4}]^2$. For families with three affected children $N_3 = \bar{\chi}^2/48\alpha^2[\theta(1 - \theta) - \frac{1}{4}]^2$. Thus, $N_2 = 3N_3$, and three times as many two-children families as three-children families are required for a specified value of χ^2.

The χ^2 test for linkage is based on computations of probabilities from tails of distributions and will therefore generate significance levels that are inevitably somewhat different from those based on the likelihood-ratio approach de-

scribed above. For practical purposes, there is sufficient agreement between results of the χ^2 and the likelihood-ratio approaches, as can be seen by comparing the required numbers of families indicated in table 1 with those calculated from the χ^2 formulas above.

Marker Locus with Many Alleles

Marker loci with many alleles are more informative than two-allele markers, and fewer families will be required on average to detect linkage. Suppose for a recessive disease, $A_1A_2 \times A_3A_4$ parents have $n = i_1 + i_2 + i_3 + i_4$ affected children as follows: i_1 children are A_1A_3, i_2 children are A_1A_4, i_3 children are A_2A_3, and i_4 children are A_2A_4. If proportion α of the families have disease due to a recessive allele linked to A_1 and A_3, and proportion $(1 - \alpha)$ of the families have disease due to recessive loci not linked to A, then the proportion of families with i_1, i_2, i_3, and i_4 children is

$$P(i_1,i_2,i_3,i_4) = \frac{n!\alpha}{i_1!i_2!i_3!i_4!}(1-\theta)^{2i_1+i_2+i_3}\theta^{i_2+i_3+2i_4} + (1-\alpha)\left(\frac{1}{2}\right)^{2n}.$$

The corresponding formula for the lod score for $A_1A_2 \times A_3A_4$ family is $\text{lod}(i_1,i_2,i_3,i_4) =$

$$\log_{10}\left\{\frac{(1+3\psi)a}{4}[(1-r)^{2i_1+i_2+i_3}r^{i_2+i_3+2i_4}] + \frac{(1-\psi)a}{4}[(1-r)^{i_1+2i_3+i_4}r^{i_1+2i_2+i_4} + (1-r)^{i_1+2i_2+i_4}r^{i_1+2i_3+i_4} + (1-r)^{i_2+i_3+2i_4}r^{2i_1+i_2+i_3}] + (1-a)\left(\frac{1}{2}\right)^{2n}\right\} - \log_{10}\left(\frac{1}{2}\right)^{2n}$$

where $\psi = 1$ if phase is known and $\psi = 0$ if phase is unknown.

The expected score per family is

$$E(L) = \sum_{i_3}\sum_{i_2}\sum_{i_1} P(i_1, i_2, i_3, i_4)\,\text{lod}(i_1, i_2, i_3, i_4).$$

As before, $N = 3/E(L)$. Tables 3 and 4 indicate the number of $A_1A_2 \times A_3A_4$ families with $n = 2$, $n = 3$, and $n = 4$ affected children required to detect linkage for various values of θ and α, phase unknown and known. The increase in amount of information using multiallelic vs. biallelic markers is apparent by comparing tables 1 and 2 with tables 3 and 4.

NUMBER OF FAMILIES NECESSARY TO DETECT HETEROGENEITY

In order to test whether more than one unlinked disease locus is present in a sample of families, we test whether the maximum lod score over all θ and α (L_{max}) is significantly greater than the maximum lod score assuming no heterogeneity and varying θ ($L_{\alpha=1}$).

TABLE 3

NO. FAMILIES REQUIRED TO DETECT LINKAGE AT 3.0, GIVEN $A_1A_2 \times A_3A_4$ PARENTS AND PHASE UNKNOWN, RECESSIVE DISEASE

	θ				
α	.001	.05	.10	.15	.20
			n = 2 affected children		
1.0	5.1	9.1	16	28	52
.9	6.7	12	19	34	65
.8	8.7	15	25	43	82
.7	12	19	32	57	107
.6	16	26	44	77	145
.5	22	37	62	110	208
.4	34	57	96	171	324
.3	59	98	168	300	573
.2	129	215	370	666	1,278
.1	493	829	1,444	2,620	5,063
			n = 3 affected children		
1.0	2.6	4.2	6.8	11	21
.9	3.2	5.2	8.4	14	25
.8	3.9	6.4	10	17	31
.7	4.9	8.1	13	22	40
.6	6.4	10	17	29	53
.5	8.6	14	24	41	75
.4	12	21	35	61	114
.3	20	33	57	102	196
.2	39	66	117	215	421
.1	128	226	415	791	1,594
			n = 4 affected children		
1.0	1.7	2.7	4.2	6.8	12
.9	2.1	3.3	5.1	8.2	14
.8	2.5	3.9	6.2	10	18
.7	3.0	4.8	7.7	13	23
.6	3.8	6.1	10	16	30
.5	4.9	8.1	13	22	41
.4	6.8	11	19	33	60
.3	10	17	30	53	101
.2	18	32	57	106	208
.1	52	96	181	356	742

Marker Locus with Two Alleles

For two-allele markers, we have already shown that it is more difficult to separate heterogeneity from recombination and, hence, test for the presence of heterogeneity. For $n = 2$ or 3 and phase unknown, the maximum likelihood approach indicates that the expected number of families for detecting heterogeneity is infinity.

For $A_1A_2 \times A_kA_k$ families with linkage phase known, or with four or more

TABLE 4

NO. FAMILIES REQUIRED TO DETECT LINKAGE AT 3.0, GIVEN $A_1A_2 \times A_3A_4$ PARENTS AND PHASE KNOWN, RECESSIVE DISEASE

	θ				
α	.001	.05	.10	.15	.20
			n = 2 affected children		
1.0	2.6	3.5	4.7	6.4	9.0
.9	3.2	4.4	5.8	8.0	11
.8	3.9	5.4	7.3	10	14
.7	4.9	6.8	9.2	12	17
.6	6.4	8.6	12	16	24
.5	8.6	12	16	23	33
.4	12	17	24	34	49
.3	20	28	39	56	83
.2	39	55	79	116	175
.1	128	186	275	415	643
			n = 3 affected children		
1.0	1.7	2.4	3.2	4.3	6.0
.9	2.1	2.9	3.8	5.2	7.4
.8	2.5	3.4	4.7	6.5	9.1
.7	3.0	4.2	5.8	8.1	12
.6	3.8	5.4	7.5	10	15
.5	4.9	7.1	10	14	21
.4	6.7	10	14	20	30
.3	10	15	22	33	49
.2	18	28	42	64	99
.1	52	82	129	207	340
			n = 4 affected children		
1.0	1.3	1.8	2.4	3.2	4.5
.9	1.5	2.1	2.9	3.9	5.5
.8	1.8	2.5	3.4	4.7	6.8
.7	2.2	3.1	4.2	5.9	8.5
.6	2.6	3.8	5.3	7.5	11
.5	3.4	4.9	7.0	10	15
.4	4.5	6.7	10	14	21
.3	6.6	10	15	22	34
.2	11	17	27	41	65
.1	28	47	75	124	210

children, there will be a unique value $\hat{r}$ at which the lod score reaches its maximum (L_{max}). It is possible to determine $\hat{r}$ as a function of θ and α by differentiating L with respect to r; that is, from equation (1),

$$\frac{\partial L}{\partial r} = \sum^{i} P(i) \frac{-a(1-r)^{n-i-1}r^{i}(n-i) + ai(1-r)^{n-i}r^{i-1}}{a(1-r)^{n-i}r^{i} + (1-a)\left(\frac{1}{2}\right)^{n}} = 0 \ .$$

When linkage phase is known, if one sets $a = 1$ but, in fact, $\alpha < 1$, the maximum lod score L_{max} occurs at

$$\hat{r} = \alpha\theta + \frac{1 - \alpha}{2} .$$

This value for $\hat{r}$ is independent of n. We understand an identical formula appears in Ott [7] and that a corresponding formula is given there for unknown phase (J. Ott, personal communication, 1985).

The P value for the test of heterogeneity is $P = 10^{-N(L_{max} - L_{\alpha=1})}$. Tables 5 and 6 indicate the number N of $A_1A_2 \times A_kA_k$ families of $n = 2$, $n = 3$, and $n = 4$ children required to detect heterogeneity based on a one-tailed test at .05, or P = .10. Therefore logP = −1.0, so that

$$N = \frac{1.0}{L_{max} - L_{\alpha=1}} . \quad (9)$$

Marker Locus with Many Alleles

For $A_1A_2 \times A_3A_4$ families with recessive disease, testing heterogeneity is easier. The test for heterogeneity is analogous to that for marker loci with two alleles, and the number of families is defined by equation (9). Tables 7 and 8 indicate the number of families with $A_1A_2 \times A_3A_4$ parents, $n = 2$, $n = 3$, and $n = 4$ children, and phase unknown or known, required to detect heterogeneity.

DISCUSSION

In the presence of genetic heterogeneity, the general shape of the curve of lod score vs. recombination fraction θ remains the same, but is displaced: the

TABLE 5

NO. FAMILIES REQUIRED TO DETECT HETEROGENEITY AT ODDS RATIO 10:1, GIVEN $A_1A_2 \times A_kA_k$ PARENTS, PHASE UNKNOWN

	θ				
α	.001	.05	.10	.15	.20
			$n = 4$		
.9	38	263	1,298	5,665	>10,000
.8	30	131	520	2,030	8,444
.7	30	104	362	1,306	5,180
.6	34	101	321	1,091	4,159
.5	41	113	334	1,080	3,990
.4	55	141	398	1,243	4,475
.3	84	207	561	1,700	6,001
.2	164	388	1,022	3,030	>10,000
.1	561	1,295	3,346	9,783	>10,000

NOTE: Dash indicates more than 10^4 families required. Heterogeneity cannot be detected for phase unknown and fewer than four children per family.

TABLE 6

NO. FAMILIES NECESSARY TO DETECT HETEROGENEITY AT ODDS RATIO 10:1, GIVEN $A_1A_2 \times A_kA_k$ PARENTS, PHASE KNOWN

	θ				
α	.001	.05	.10	.15	.20
			$n = 2$		
.9	50	142	368	911	2,253
.8	37	80	172	374	844
.7	36	68	131	263	559
.6	39	67	122	233	475
.5	45	75	131	241	476
.4	59	95	160	286	551
.3	89	138	228	399	756
.2	168	257	420	721	1,349
.1	562	852	1,370	2,345	4,356
			$n = 3$		
.9	18	54	130	306	738
.8	14	30	64	133	292
.7	13	25	48	95	197
.6	13	24	44	84	168
.5	15	26	47	85	167
.4	20	32	55	99	190
.3	28	45	76	134	255
.2	50	80	133	233	442
.1	157	247	410	721	1,370
			$n = 4$		
.9	10	30	69	156	367
.8	7.6	17	35	70	151
.7	7.0	14	26	51	103
.6	7.2	13	24	45	88
.5	8.1	14	25	45	88
.4	10	17	29	52	99
.3	14	23	39	69	130
.2	24	38	65	115	219
.1	68	110	188	337	651

average lod score per family may decrease and the peak lod score is found at an r value higher than the true θ, thus imitating looser linkage. Therefore, the sensitivity of the test for linkage is decreased, and detection of linkage in the presence of heterogeneity will require more families. We have estimated the expected number of families required to detect linkage given heterogeneity compared to the expected number required in the absence of heterogeneity. The types of matings considered are the common ones: for a dominant disease, $DdA_1A_2 \times ddA_kA_k$, including all children; and for a recessive disease, $DdA_1A_2 \times DdA_kA_k$, and $DdA_1A_2 \times DdA_3A_4$ including only affected children.

The number of families required to detect linkage increases as the proportion

TABLE 7

NO. FAMILIES REQUIRED TO DETECT HETEROGENEITY AT ODDS RATIO 10:1, GIVEN $A_1A_2 \times A_3A_4$ PARENTS AND PHASE UNKNOWN, RECESSIVE DISEASE

	θ				
α	.001	.05	.10	.15	.20
			n = 2 affected children		
.9	50	335	1,564	6,449	>10,000
.8	38	154	605	2,258	9,025
.7	36	122	410	1,423	5,464
.6	39	114	354	1,165	4,335
.5	46	124	359	1,134	4,111
.4	60	151	419	1,285	4,567
.3	89	216	579	1,736	6,070
.2	169	397	1,038	3,062	>10,000
.1	566	1,303	3,360	9,799	>10,000
			n = 3 affected children		
.9	18	121	511	1,849	6,680
.8	13	53	184	606	2,095
.7	12	38	115	356	1,192
.6	12	33	92	271	884
.5	13	32	85	243	780
.4	15	36	91	252	800
.3	20	46	112	307	973
.2	33	72	176	479	1,533
.1	90	196	478	1,324	4,334
			n = 4 affected children		
.9	7.4	37	137	501	1,913
.8	5.3	19	58	185	639
.7	4.8	15	40	117	379
.6	4.8	13	34	93	288
.5	5.2	13	32	85	257
.4	6.1	14	34	88	262
.3	7.8	18	42	105	311
.2	12	27	61	156	465
.1	28	62	146	384	1,193

α of families with disease linked to the marker decreases. Just to give an indication of the order of magnitude, the expected number of families required to detect linkage is three to five times greater if α = .50 than if α = 1.0. This is a substantial increase, but it may still be possible to detect linkage especially if it is reasonably tight. There are obviously three ways of improving the chances of detecting linkage: (1) include families with a larger number of informative children (more likely for dominant diseases, since affected and unaffected children are equally useful); (2) use collateral relatives to establish the phase (also more likely to be possible for dominant diseases); and (3) for recessive diseases, find more polymorphism at the marker. Frequently, heterozygosity of a marker can

TABLE 8

NO. FAMILIES REQUIRED TO DETECT HETEROGENEITY AT ODDS RATIO 10:1, GIVEN $A_1A_2 \times A_3A_4$ PARENTS AND PHASE KNOWN, RECESSIVE DISEASE

	θ				
α	.001	.05	.10	.15	.20
			n = 2 affected children		
.9	10	30	69	156	367
.8	7.6	17	35	70	151
.7	7.0	14	26	51	103
.6	7.2	13	24	45	88
.5	8.1	14	25	45	88
.4	10	17	29	52	99
.3	14	23	39	69	130
.2	24	38	65	115	219
.1	68	110	187	336	651
			n = 3 affected children		
.9	5.2	14	30	65	148
.8	3.7	8.2	16	31	64
.7	3.3	6.7	12	23	45
.6	3.4	6.3	11	20	39
.5	3.7	6.6	11	20	38
.4	4.3	7.6	13	23	42
.3	5.7	10	17	29	55
.2	9.0	15	26	46	88
.1	23	38	67	123	242
			n = 4 affected children		
.9	3.4	8.7	18	37	81
.8	2.4	5.1	10	18	37
.7	2.1	4.2	7.4	13	26
.6	2.1	3.9	6.7	12	22
.5	2.3	4.0	6.8	12	22
.4	2.6	4.6	7.7	13	24
.3	3.3	5.8	10	17	31
.2	5.0	8.7	15	26	48
.1	11	20	35	63	125

be increased by testing with other restriction enzymes. Tables 1–4 give numerical values for evaluating the advantage to be gained with various strategies.

In general, heterogeneity is most easily detected when α is about .50. Not surprisingly, very high and very low heterogeneity are difficult to detect. At $\alpha = .50$, the sample size necessary to detect heterogeneity is comparable to that necessary for detecting linkage, especially if recombination is low. Detecting heterogeneity is very difficult when phase is unknown, for either recessive or dominant diseases. With two-allele markers and unknown phase, heterogeneity and linkage can be separately estimated if $n \geq 4$, but there are not many families with recessive disease with four or more affected children. The

situation is better for dominant diseases, since all children are equally informative. For recessive diseases, four-allele matings $A_1A_2 \times A_3A_4$ enable the detection of heterogeneity even if phase is unknown.

In summary, heterogeneity increases the difficulty of detecting linkage, although both linkage and heterogeneity can be estimated under appropriate conditions. For dominant, fully penetrant diseases, the task is easier than for recessives because phase can be ascertained and because unaffected children are as informative as affected ones. For recessive diseases, the geneticist's best allies are highly polymorphic markers.

ACKNOWLEDGMENTS

Dr. J. Ott gave helpful suggestions for this manuscript and told us of the similarities between our treatment and that in his recent book, which we had not had a chance to examine properly at the time of revision.

REFERENCES

1. Morton NE: Sequential tests for the detection of linkage. *Am J Hum Genet* 7:277–318, 1955
2. Smith CAB: Testing for heterogeneity of recombination fraction values in human genetics. *Ann Hum Genet* 27:175–182, 1963
3. Ott J: Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann Hum Genet* 40:443–454, 1977
4. Ott J: Linkage analysis and family classification under heterogeneity. *Ann Hum Genet* 47:311–320, 1983
5. Risch N, Baron M: X-linkage and genetic heterogeneity in bipolar-related major affective disorder: reanalysis of linkage data. *Ann Hum Genet* 46:153–166, 1982
6. Hodge SE, Anderson CE, Neiswanger K, Sparkes RS, Rimoin DL: The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. *Am J Hum Genet* 35:1139–1155, 1983
7. Ott J: *Analysis of Human Genetic Linkage*. Baltimore, Johns Hopkins Univ. Press, 1985