Am J Hum Genet 39:513–527, 1986

Estimating the Power of a Proposed Linkage Study: A Practical Computer Simulation Approach

MICHAEL BOEHNKE[1]

SUMMARY

I describe a simulation method to estimate the power to detect linkage given a set of pedigrees of known structure and for which family history data may be available. This method can be applied to autosomal and X-linked dominant diseases; depending on the pedigrees under consideration, it will often be applicable for autosomal and X-linked recessive diseases. This power calculation can most usefully be undertaken after family history data are gathered, but prior to examination and testing of pedigree members to obtain marker information. Of key importance, the power calculation is straightforward to carry out and not too time-consuming; it is practical even on a microcomputer. The result of the power calculation is an objective answer to the question: Will my families be sufficient to demonstrate linkage?

INTRODUCTION

There are two requirements to map a Mendelian disease by standard linkage methods: a polymorphic genetic marker linked to the disease locus and a sufficient set of informative pedigrees. Until recently, the human genetic map was quite sparse, with relatively few useful markers covering only a small portion of the genome. Thus, attention was focused on the existence of a linked marker, and the prior probability of linkage was of primary concern [1, 2].

Given a spanning set of 150–200 markers spaced at 20-cM intervals throughout the genome, any disease gene should be within at most 10 cM of a known marker [3]. While many more markers will be required before such a spanning

Received March 25, 1986.

This research was supported by grants P01-CA-26803 and R01-HL-24489 from the National Institutes of Health.

[1] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109.

set is obtained [4, 5], the current explosive increase in the number of available restriction fragment length polymorphisms (RFLPs) means that a spanning set soon will exist. With the requirement for a linked marker satisfied, the availability of sufficiently many informative pedigrees will take on primary importance. Thus, it becomes critical to determine whether a given set of pedigrees are likely to be sufficiently informative to map a genetic disease *assuming* a linked marker exists.

Here, I describe a simulation approach to estimate the power to demonstrate linkage given a set of pedigrees of known structure. This approach is most usefully undertaken after family history data have been gathered, but before families are actually seen, blood drawn, and lab work carried out. At the family history stage, relatively little effort has been expended, but the pedigree structure and the disease phenotypes of some of the pedigree members are known.

The results of this simulation approach provide objective a priori evidence regarding the statistical power of a proposed study to identify a linked marker. In addition, they suggest which pedigrees are likely to be most informative. Finally, carrying out the simulation at several true recombination fractions can suggest an appropriate choice of marker spacing for a linkage study. If the available pedigrees are shown to have a high probability of demonstrating linkage given a spanning set of markers equally spaced at 20-cM intervals, such a spacing would be suggested. If not, either more pedigrees or a more densely spaced spanning set would be required. Carrying out the simulation for unlinked loci can also provide an estimate of the distance from an unlinked marker that is likely to be excluded and the probability of incorrectly concluding linkage to an unlinked marker.

METHODS

I begin this section by outlining a strategy to evaluate the linkage information in a pedigree and a set of pedigrees; each step of the strategy is then discussed in greater detail.

Brief Outline

A. For each of the available pedigrees and at each of several true recombination fractions θ_j (for example, $\theta = 0, .05, .10, .15, .20$, and $.50$):

1. *Simulation of pedigrees*. Simulate cosegregation of the disease locus and a marker locus in N copies (for example, $N = 1{,}000$) of the pedigree assuming a true recombination fraction θ_j. Individuals of known disease phenotype are assigned disease genotypes compatible with their disease phenotypes.

2. *Calculation of lod scores*. Calculate lod scores $Z(r_i; \theta_j)$ for the simulated pedigrees at each of several test recombination fractions r_i. The lod score $Z(r; \theta)$ is the logarithm base 10 of the pedigree likelihood assuming a recombination fraction r ($.00 \leq r \leq .50$) divided by the pedigree likelihood assuming free recombination ($r = .50$). The notational dependence of the lod score on θ will be used here to indicate that θ is the true recombination fraction under which the pedigree data were simulated.

3. *Linkage information criteria for each pedigree*. For each test recombina-

tion fraction r_i, tabulate the lod scores $Z(r_i; \theta_j)$ and the approximate maximum lod scores $Z^*(\theta_j) = \max_i Z(r_i; \theta_j)$ for the N simulated pedigrees. From these lod score distributions, estimate the linkage information criteria for the pedigree: the probability of a maximum lod score greater than some constant c, and the expected maximum lod score.

B. For the set of pedigrees:

4. *Joint linkage information in a set of pedigrees*. Estimate the joint linkage information provided by all the available pedigrees: the probability of a maximum summed lod score for the pedigrees greater than some constant c, or of a maximum lod score for at least one pedigree greater than c, and the expected maximum summed lod score.

1. *Simulation of Pedigrees*

For a simulation approach to this problem to be practical, the simulation procedure must efficiently assign pedigree two-locus genotype vectors $\mathbf{g} = (g_1, g_2, \ldots, g_p)$ to the p members of the pedigree. In addition, the simulation procedure must avoid any systematic bias in the pedigree genotype vectors it assigns; that is, the probability the simulation assigns a two-locus genotype vector $\mathbf{g}$ to a pedigree with disease phenotypes $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ should be equal to $P(\mathbf{g}|\mathbf{x})$, the conditional probability of $\mathbf{g}$ given $\mathbf{x}$ (see below).

Once genotypes are simulated for each member of a pedigree, corresponding disease and marker phenotypes are recorded in a pedigree file to be used in the subsequent linkage analysis. Pedigree members who are expected to be available for sampling in the linkage study are assigned the disease and marker phenotypes corresponding to their simulated genotypes. Pedigree members who are deceased or for some other reason unlikely to be available are assigned unknown marker phenotypes.

Disease phenotypes unknown. If disease phenotypes are unknown for all pedigree members, the requirement for efficient, unbiased simulation of genotype vectors is easy to satisfy for any Mendelian trait. In this case,

$$P(\mathbf{g}|\mathbf{x}) = P(\mathbf{g}) = \prod_{k \in O} P(g_k) \prod_{l \in D} P(g_l|g_{l_1}, g_{l_2}) , \tag{1}$$

where k runs over all originals ("O" indicates the set of originals) in the pedigree, l runs over all descendants ("D" indicates the set of descendants), and l_1 and l_2 are the parents of l. Equation (1) immediately suggests an approach to simulate genotype vectors. First, genotypes are simulated for each pedigree original according to the prior probabilities for the genotypes given gene frequencies and the assumptions of Hardy-Weinberg and linkage equilibrium. Second, genotypes are simulated for descendants based on the genotypes of the parents according to transmission probabilities assuming Mendelian segregation and a recombination fraction θ assumed equal in males and females.

Disease phenotypes known. When disease phenotypes are known for some of the pedigree members, as will be true after family history data are gathered,

simulating genotype vectors for the pedigree in an unbiased manner can be much more difficult. Efficient simulation requires that the conditional probability $P(\mathbf{g}|\mathbf{x})$ again factor into terms each dependent on the genotypes of one or a few pedigree members. In analogy to the phenotype-unknown case, if the conditional probability can be expressed as

$$P(\mathbf{g}|\mathbf{x}) = \prod_{k \epsilon O} P(g_k|x_k) \prod_{l \epsilon D} P(g_l|g_{l_1}, g_{l_2}, x_l) , \qquad (2)$$

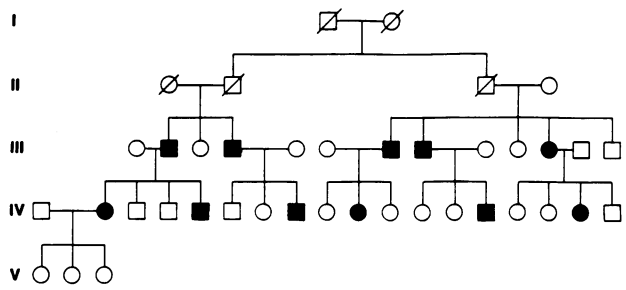
unbiased simulation of genotype vectors can be carried out efficiently. Sufficient (although not necessary) conditions for equation (2) to hold are: (a) known disease phenotype (effectively) implies known disease genotype, and (b) disease genotypes are known or can be inferred for the parents of all descendants with known disease genotype (see APPENDIX A). These requirements hold for an autosomal or X-linked dominant disease; depending on the pedigrees under consideration, they often hold for autosomal and X-linked recessive diseases as well. The reason is that such diseases are rare, so that certain disease genotypes are very unlikely and may reasonably be ignored.

As an example, consider two pedigrees in which a rare autosomal dominant disease is segregating (fig. 1). For pedigree A, family history data provide disease phenotypes for all pedigree members except I-1, I-2, II-1, II-2, and II-3. Since the disease is autosomal dominant, unaffecteds are homozygous normal *dd*. Since the disease is rare, affected individuals are almost certainly heterozygous *Dd*, II-2 and II-3 were almost certainly *Dd*, and II-1 was almost certainly *dd*. Finally, either I-1 or I-2 was affected; these events are equally likely. Assign *Dd* to I-1 and *dd* to I-2; the assumption of equal male and female recombination fractions means that this choice will make no difference in the results. For pedigree B, the same approach assigns *Dd* and *dd* to I-1 and I-2, and *Dd* and *dd* to all other affecteds and unaffecteds, respectively. Thus, all pedigree members can be assigned a disease genotype.

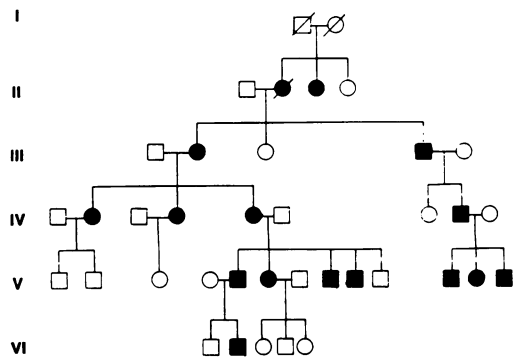
If disease penetrance is incomplete, the conditional probability $P(\mathbf{g}|\mathbf{x})$ often does not factor into a product of simple terms, since several disease genotypes may be possible for some unaffected pedigree members. Nonetheless, if disease penetrance is complete or nearly so by a given age, the approach outlined above may still be used. The simulation can be carried out first with the younger unaffecteds excluded and second with pedigrees and phenotypes as observed, resulting in lower and upper bounds on the linkage information. This approach allows us to again assume that known disease phenotype specifies known disease genotype.

2. *Calculation of Lod Scores*

For each replicate pedigree and each true recombination fraction θ, lod scores are calculated at several test recombination fractions r_i, using one of the available linkage analysis programs. These include LIPED [6, 7], PAP [8], and LINKAGE [9]. If $\theta = 0$, the maximum lod score must occur at $r = 0$, and so only



Pedigree A



Pedigree B

FIG. 1.—Two pedigrees in which an autosomal dominant disease is segregating

one test recombination fraction is required. If $0 < \theta < .50$, the r_i should include values bracketing the true recombination fraction θ. For example, if $\theta = .10$, one might consider $r = .01, .05, .10, .15$, and $.20$. If $\theta = .50$, a similar choice of r_i will suggest to what distance from an unlinked marker linkage might be excluded. Just as efficient simulation of pedigrees is required for this power calculation to be practical, so too is efficient lod score calculation essential.

3. *Linkage Information Criteria*

Information provided by a single linked marker. Having calculated the pedigree lod scores, a variety of criteria can be used as the information for linkage in a pedigree. Here, I consider two: the probability that the (approximate) maximum pedigree lod score $Z^*(\theta) = \max_i Z(r_i; \theta)$ is greater than some constant c, and the expected value of the maximum pedigree lod score. The sample proportion $\hat{p}$ of maximum lod scores greater than c provides an estimate of the probability of a maximum lod score greater than c; $[\hat{p}(1 - \hat{p})/n]^{1/2}$ is the standard error of the estimator. The sample mean of the maximum lod scores provides an estimate of the expected maximum lod score.

For $\theta < .50$, $P(Z^*(\theta) \geq 3.0)$ represents the statistical power of the pedigree to

reject the hypothesis of no linkage when the two loci are linked. For $\theta = .50$, $P(Z^*(\theta) \geq 3.0)$ is the probability of erroneously concluding linkage to a single unlinked marker, that is, of making a type I error. If m markers are to be tested, an estimate of the overall type I error probability taking into account all m tests is $1 - P[Z^*(.50) < 3.0]^m$, assuming independence. Analogous calculations can be made for a set of pedigrees.

Simulating pedigrees at several true recombination fractions θ_j suggests the effect of the distance to a marker on the linkage information. For linked loci ($\theta < .50$), calculating $Z(r_i; \theta)$ at several r_i allows for the possibility that the maximum lod score for a given pedigree may occur at a test recombination fraction $r \neq \theta$. This is important since the largest lod scores result in pedigrees where (perhaps by chance) little or no recombination occurred. Thus, the largest pedigree lod scores will often occur at a test recombination fraction $r < \theta$. For unlinked loci ($\theta = .50$), calculating $Z(r_i; \theta)$ at several r_i suggests the length of the region that might be excluded by testing for linkage to an unlinked marker.

Information provided by a spanning set of markers. If a spanning set of markers equally spaced at d cM intervals is to be used, a disease locus will be within t and $(d - t)$ cM of the two closest markers ($0 \leq t \leq d/2$), as well as $(d + t)$ and $(2d - t)$ cM of the next closest markers, and so forth. Thus, in calculating the probability of achieving a maximum lod score greater than c when a spanning set of markers is to be used, we should consider not only the closest linked marker but also neighboring marker(s). To keep the number of true recombination fractions θ in the simulation manageably small, I consider only those markers no more than d cM from the disease locus; that is, the two closest markers when $t > 0$, and the three closest when $t = 0$. The result is an underestimate of the power; however, unless d is rather small, these closest markers provide most of the linkage information, and the underestimation should be minor. If d is small, what follows can be modified to take into account additional linked markers.

Let $\hat{M}_d(t)$ be the maximum lod score to any linked marker at most d cM away, assuming a spanning set of markers evenly spaced at d cM intervals and conditional on t cM being the distance to the nearest marker. Let $M_d^*(t)$ be the corresponding maximum for the test recombination fractions r_i considered. Then if θ_t is the recombination fraction corresponding to a map distance of t cM, the probability that $\hat{M}_d(t)$ should be greater than c may be estimated by

$$P[M_d^*(t) \geq c] = \begin{cases} 1 - P[Z^*(\theta_t) < c]P[Z^*(\theta_{d-t}) < c] & t > 0 \\ 1 - P[Z^*(0) < c]^2 P[Z^*(\theta_d) < c] & t = 0. \end{cases} \quad (3)$$

The standard error of this estimator is given in APPENDIX B.

Assuming that the distance t from the disease locus to the nearest marker is uniformly distributed on the interval 0 to $d/2$, the (unconditional) probability that the maximum lod score $\hat{M}_d$ to any linked marker up to d cM away should

be greater than c is $P[\hat{M}_d \geq c] = 2/d \int_0^{d/2} P[\hat{M}_d(t) \geq c]\, dt$. Applying Simpson's rule for numerical integration [10] and approximating $\hat{M}_d(t)$ by $M_d^*(t)$,

$$\begin{aligned} P(\hat{M}_d \geq c) &\approx P(M_d^* \geq c) \\ &\approx \tfrac{1}{6}\{P[M_d^*(0) \geq c] + 4P[M_d^*(d/4) \geq c] + P[M_d^*(d/2) \geq c]\}. \end{aligned} \tag{4}$$

The standard error for this estimator is easily calculated since if $X_1, X_2, \ldots$ are independent random variables and $c_1, c_2, \ldots$ are constants, then $\mathrm{Var}(\Sigma_k c_k X_k) = \Sigma_k c_k^2 \mathrm{Var}(X_k)$.

For small to moderate map distances, say $t \leq 25$ cM, $\theta_t \approx t/100$ for mapping functions θ_t taking into account inference. For larger t, one might want to explicitly evaluate the mapping function θ_t to choose the appropriate recombination fractions to simulate.

For example, suppose $d = 20$ cM. Then $\theta_t \approx t/100$ for all $t \leq d$. To estimate the power to detect a linked marker: First, estimate probabilities of lod scores greater than c for pedigrees with true recombination fractions $\theta = 0$ and .20; $\theta = .05$ and .15; and $\theta = .10$. Second, substitute these results in equation (3) to estimate the conditional probability of a maximum lod score greater than c to any linked marker assuming distances to the nearest marker of 0, 5, or 10 cM, respectively. Third, substitute the estimates of $M_{20}^*(0)$, $M_{20}^*(5)$, and $M_{20}^*(10)$ into equation (4) to estimate the unconditional probability of a maximum lod score greater than c to any linked marker.

If the estimated power for a spanning set at 20 cM intervals is insufficient, more pedigrees might be sought. Alternatively, one might consider an analysis with a denser marker spacing; for example, with $d = 10$ cM. The power when $d = 10$ cM can immediately be approximated by $\frac{1}{2}[M_{10}^*(0) + M_{10}^*(5)]$; this estimate can be calculated directly from the simulation results for $\theta = 0$, .05, and .10 already undertaken. A more accurate approximation is obtained by additional simulations with $\theta = .025$ and .075 and application of Simpson's rule [10].

If the estimated power is very close to one, a less dense spanning set might be considered. Again, a portion of the results from the $d = 20$ cM simulations should be reused to keep the additional simulation work to a minimum.

4. *Joint Linkage Information in a Set of Pedigrees*

Given a set of pedigrees with family history data, carrying out the above procedure for each available pedigree suggests which pedigrees are likely to provide the most linkage information. In addition, we can estimate the total linkage information for the set of pedigrees.

First, the probability that the maximum summed lod score should be greater than some constant c can be estimated. Since a large number of replicates N will be desirable, a reasonable estimate of the probability of a maximum summed lod score greater than c is obtained by summing the lod scores $Z(r_i; \theta)$ for the nth replicate of each pedigree ($1 \leq n \leq N$) and calculating the proportion of

these maximum summed values greater than c. The alternative approach of convolving the distributions for the different pedigrees and finding their maximum would be much more difficult computationally and would quickly become impractical for even a modest number of pedigrees and replications. The procedure outlined above for the lod scores for single pedigrees can then be repeated for the summed lod scores.

Second, the probability that at least one of the available pedigrees will by itself have a maximum lod score greater than c can be estimated. This probability is of particular interest when genetic heterogeneity is a concern; given genetic heterogeneity, summing the lod scores for several pedigrees is inappropriate. Assuming markers evenly spaced at d cM intervals, the probability that at least one pedigree will have a lod score greater than c to at least one of the linked markers is estimated by

$$1 - \prod_k P({}_kM_d{}^* < c) \,. \tag{5}$$

Here, ${}_kM_d{}^*$ is the approximate maximum lod score to any linked marker for the kth observed pedigree. The standard error for this estimator is given in APPENDIX B.

Third, the expected maximum summed lod score may be calculated.

APPLICATION

As an application of this method, consider the family history data illustrated in figure 1. A rare, fully penetrant autosomal dominant disease is segregating in these pedigrees. Genotype inferencing for the pedigrees was described above.

I simulated cosegregation of the dominant disease and a linked marker in $N = 1{,}000$ replicates each of pedigrees A and B at true recombination fractions $\theta = 0, .05, .10, .15, .20$, and $.50$. Table 1 lists the test recombination fraction r_i at which lod scores were calculated for each true recombination fraction θ. Simulations were carried out using the FORTRAN program SIMLINK which I wrote for this purpose. Lod scores were calculated using LIPED [6, 7]. In each analysis, I assumed a two-allele, codominant marker with allele frequencies .50.

Table 2 presents the mean maximum lod scores and their standard errors for

TABLE 1

TEST RECOMBINATION FRACTIONS r_i

TRUE RECOMBINATION FRACTION θ	TEST RECOMBINATION FRACTIONS				
	r_1	r_2	r_3	r_4	r_5
.00	.00	...	...	...	...
.05	.01	.03	.05	.10	.20
.10	.01	.05	.10	.15	.20
.15	.01	.05	.10	.15	.20
.20	.01	.05	.10	.20	.30
.50	.01	.05	.10	.15	.20

TABLE 2

MEAN MAXIMUM LOD SCORE ± STANDARD ERROR FOR A DOMINANT DISEASE AND A LINKED CODOMINANT MARKER

TRUE RECOMBINATION FRACTION θ	PEDIGREE A	PEDIGREE B	PEDIGREE A + B*
.00	2.84 ± .04	2.57 ± .04	5.41 ± .06
.05	2.07 ± .04	1.92 ± .04	3.84 ± .06
.10	1.52 ± .04	1.45 ± .03	2.75 ± .05
.15	1.05 ± .03	1.02 ± .03	1.89 ± .04
.20	0.70 ± .02	0.77 ± .02	1.28 ± .03

NOTE: Means and standard errors are based for each θ on N = 1,000 simulated pedigrees assuming a two-allele codominant marker with allele frequencies .50.

* Maximum summed lod score for pedigrees A and B together.

each $\theta < .50$. As expected, the mean maximum lod score decreased as the true recombination fraction θ increased for both pedigrees. The average maximum summed lod score ranged from 5.41 for $\theta = 0$ to 1.28 for $\theta = .20$. For unlinked loci ($\theta = .50$), the means of the summed lod scores were uniformly negative (data not shown). For $r \leq .10$, the mean summed lod score was less than -2.0, the value customarily accepted as excluding linkage.

Table 3 reports the estimated probabilities of maximum lod scores greater than 2.0 and greater than 3.0 assuming a single linked marker. Probabilities of maximum lod scores greater than 2.0 or 3.0 also decreased with increasing θ. The probability that the maximum summed lod score for A and B together ex-

TABLE 3

PROBABILITIES OF A MAXIMUM LOD SCORE GREATER THAN 2.0 OR 3.0 ± STANDARD ERROR FOR A DOMINANT DISEASE AND A LINKED CODOMINANT MARKER

TRUE RECOMBINATION FRACTION θ	c	PEDIGREE A	PEDIGREE B	PEDIGREE A + B*	PEDIGREE A or B†
.00	2.0	.71 ± .01	.66 ± .01	.97 ± .01	.90 ± .01
	3.0	.44 ± .02	.34 ± .02	.91 ± .01	.64 ± .01
.05	2.0	.49 ± .02	.43 ± .02	.85 ± .01	.71 ± .01
	3.0	.24 ± .01	.19 ± .01	.65 ± .02	.39 ± .02
.10	2.0	.30 ± .01	.27 ± .01	.64 ± .02	.49 ± .01
	3.0	.11 ± .01	.10 ± .01	.39 ± .02	.20 ± .02
.15	2.0	.18 ± .01	.14 ± .01	.41 ± .02	.29 ± .01
	3.0	.06 ± .01	.04 ± .01	.19 ± .01	.10 ± .01
.20	2.0	.07 ± .01	.08 ± .01	.21 ± .01	.15 ± .01
	3.0	.02 ± .004	.01 ± .004	.08 ± .01	.03 ± .01

NOTE: Means and standard errors are based for each θ on N = 1,000 simulated pedigrees assuming a two-allele codominant marker with allele frequencies .50.

* Probabilities that the maximum summed lod score for the two pedigrees should be greater than c.

† Probabilities that the maximum lod score for at least one pedigree should be greater than c.

TABLE 4

PROBABILITIES OF A MAXIMUM LOD SCORE GREATER THAN 2.0 OR 3.0 ± STANDARD ERROR FOR A DOMINANT DISEASE AND A SPANNING SET OF CODOMINANT MARKERS SPACED AT d = 20 cM INTERVALS

		PEDIGREE			
t(cM)	c	A	B	A + B*	A or B†
		A. Probabilities $M_d^*(t)$ conditional on the distance t to the nearest marker			
0	2.0	.75 ± .01	.71 ± .01	.98 ± .003	.93 ± .01
	3.0	.47 ± .02	.36 ± .02	.92 ± .01	.66 ± .01
5	2.0	.58 ± .01	.51 ± .01	.91 ± .01	.80 ± .01
	3.0	.29 ± .01	.23 ± .01	.72 ± .01	.45 ± .01
10	2.0	.52 ± .02	.47 ± .02	.87 ± .01	.74 ± .01
	3.0	.21 ± .02	.18 ± .02	.63 ± .02	.36 ± .02
		B. Unconditional probabilities M_d^* assuming a marker spacing of d = 20 cM			
	2.0	.60 ± .01	.54 ± .01	.92 ± .01	.81 ± .01
	3.0	.30 ± .01	.24 ± .01	.74 ± .01	.47 ± .01

NOTE: Means and standard errors are based for each θ on N = 1,000 simulated pedigrees assuming two-allele codominant markers with allele frequencies .50.

* Probabilities that the maximum summed lod score for the two pedigrees should be greater than c.

† Probabilities that the maximum lod score for at least one pedigree should be greater than c.

ceeded 2.0 or 3.0 ranged from .97 and .91 for θ = 0, to .21 and .08 for θ = .20. The probability that at least one of the pedigrees had a maximum lod score greater than 2.0 or 3.0 ranged from .90 and .64 for θ = 0 to .15 and .03 for θ = .20. The maximal lod scores often occurred at a test recombination fraction $r < \theta$ (data not shown). For example, for θ = .05, the probability of a lod score greater than 3.0 was greater at r = .01 (.225 for pedigree A, .184 for pedigree B) than at r = .05 (.192 for A, .145 for B).

Table 4 presents the probabilities of achieving a maximum lod score greater than 2.0 or 3.0 to at least one linked marker up to 20 cM distant. When the nearest marker is t = 0 cM away, these probabilities are only slightly elevated above those for a single linked marker with θ = 0 (table 3), suggesting that the additional two markers 20 cM away provided little additional power. For t = 5 cM and particularly for t = 10 cM, the improvement in power by considering the two closest markers is substantial. For example, a single linked marker 10 cM distant gave probabilities .64 and .39 for maximum summed lod scores greater than 2.0 or 3.0, respectively (table 3); considering two such markers increased these probabilities to .87 and .63 (table 4).

Using Simpson's Rule [10] to summarize these values in a single estimate of the power for a spanning set of markers at 20 cM intervals, I found that together the pedigrees gave probabilities .92 and .74 of a maximum summed lod score greater than 2.0 and 3.0, respectively. Further, the probabilities that at least one of the pedigrees should by itself have a lod score greater than 2.0 or 3.0 was .81 or .47, respectively. Thus, a linkage study based on those two pedigrees and

a spanning set of markers equally spaced at 20-cM intervals would have about a 74% chance of demonstrating linkage. If heterogeneity was suspected, so that the lod scores for the pedigrees could not be summed, this power estimate would be reduced to 47%.

By estimating the probability of a lod score greater than 2.0 or 3.0 when $\theta = .50$, I also estimated the probability of incorrectly inferring linkage to an unlinked marker, that is, of making a type I error. Among the 1,000 replicates each of pedigrees A and B, only two replicates resulted in a lod score as large as 2.0; no lod score was greater than 2.37. Considering the summed lod scores; only for one pedigree pair was a lod score greater than 2.0 achieved; the summed value for that pair had a maximum value of 2.25. Thus, even if many markers are tested for linkage in these pedigrees, the probability of concluding linkage when it is not present is small.

DISCUSSION

The simulation approach suggested in this paper is the analog of power calculations used in more standard statistical problems. In linkage analysis, one tests the null hypothesis of no linkage ($\theta = .50$) against the alternative of linkage ($\theta < .50$). Approximate control of type I error (that is, concluding linkage when $\theta = .50$) is achieved by appropriate choices of the value at which one concludes linkage [11, 12]. However, power considerations for linkage studies have received relatively little consideration [13–15], and most of that has been restricted to nuclear family data. Clearly the reason for this is the variability of possible pedigree structure and phenotypes and the dependence of the lod score on pedigree structure, phenotypes, and recombination fraction.

Skolnick et al. [13] and Elston and Bonney [15] calculated the expected lod score for offspring of phase-known, double-backcross matings and from it the expected numbers of offspring to demonstrate linkage as a function of the recombination fraction. Further, Skolnick et al. [13] calculated the expected lod scores for sibships of various sizes and the expected numbers of sibships required to demonstrate linkage. While their results are instructive, they do not provide the expected lod score for a particular pedigree, since it is usually not clear how many fully informative offspring are present in a given pedigree and since related sibships analyzed as intact pedigrees provide more information than the same number of sibships analyzed separately [14]. Further, they give no direct estimate of power.

An alternative approach to linkage information based on classical likelihood methods was suggested by Ott [16]. He calculated the Fisher information (see, for example, [17]) for the recombination fraction, given sibship data and several parental mating types. While this approach could be generalized to pedigrees, it does not directly answer our question of primary interest, namely: Is there sufficient family data to establish linkage if a linked marker exists?

In contrast, the simulation approach described here provides a direct estimate of the power to detect linkage. Further, the method is practical. For pedigrees A and B, simulating 1,000 replicates, calculating 1,000 sets of lod scores, and reading and manipulating the lod scores to estimate the linkage

information criteria required about 60 min elapsed time on an IBM-AT microcomputer, or a total of about 12 hrs elapsed time for the entire analysis reported. The time and expense of a large-scale linkage study, particularly one in which cell lines are to be grown for each individual and a significant portion of the genome spanned, easily justify the effort required to carry out such an analysis.

The power calculation suggested here does require certain assumptions. First, I assume that family history data are accurate. Obviously, errors in reporting may occur; however, any effect on the linkage information is likely to be small. Second, one must choose the type of marker to simulate. The choice of a two-allele, codominant marker with equal allele frequencies represents a compromise between highly polymorphic RFLPs with many alleles and two-allele markers that barely satisfy the frequency definition of a polymorphism. As the spanning set of markers for the human genome is constructed, preference will be given to markers that are highly polymorphic, so that this choice of markers to simulate should not be overly optimistic. Assuming a two-allele marker rather than a multiple-allele marker saves simulation time and computation time. Third, I have assumed that because genetic diseases are rare, certain genotypes (for example, homozygous *DD* for an autosomal dominant disease) do not occur. This may occasionally be wrong. Fourth, male and female recombination fractions are likely different, although here I have assumed they are equal. In fact, such simplifications and minor errors really are not of great concern. In this analysis, as in any power calculation, one seeks an approximate answer to our information questions. Of more fundamental importance, to determine the power provided by a set of pedigrees, one must assume a mode of inheritance for the disease. If mode of inheritance is uncertain, any conclusions based on the linkage information analysis are contingent on the mode of inheritance assumption being correct.

CONCLUSION

In this paper, I have described a simulation approach to estimate the information for linkage in a set of pedigrees. This approach is practical: it requires a limited amount of time, can be carried out prior to the bulk of the effort of a linkage study, and provides objective evidence on the chance that the study will be successful. For an investigator deciding whether his or her time would be well spent on a particular linkage study, this sort of evaluation should be of great interest.

COMPUTER PROGRAMS

Two FORTRAN programs were written to carry out the analysis described: SIMLINK simulates the replicate pedigrees and LODSTAT reads the lod score file and calculates the linkage information statistics. Source code for both programs, together with documentation and sample analyses, are available from the author at a nominal charge to cover the cost of a floppy disk and mailing. If

a linkage analysis program other than LIPED is used, the subroutine in LODSTAT which reads the lod score files will require modification.

ACKNOWLEDGMENTS

I thank Mr. Rork Kuick and Drs. Susan Hodge, Peter Smouse, Kenneth Lange, Patricia Moll, and Francis Collins for their helpful comments and criticisms of a previous version of this paper. Mrs. Patrice Somerville typed the manuscript quickly and accurately.

APPENDIX A

SIMULATING LINKAGE IN A PEDIGREE WITH KNOWN DISEASE GENOTYPES

The efficient simulation of linkage in pedigrees is possible if the conditional probability of the two-locus genotype vector $\mathbf{g} = (g_1, g_2, \ldots, g_p)$ given the disease locus phenotype vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ factors as in equation (2). I now demonstrate that this factorization holds if (a) known disease phenotype implies known disease genotype and (b) disease genotypes are known or can be inferred for the parents of all descendants with known disease genotype.

Let $\mathbf{h} = (h_1, h_2, \ldots, h_p)$ be the vector of disease locus genotypes. Write $\mathbf{g} = (\mathbf{g}', \mathbf{g}'')$ and $\mathbf{h} = (\mathbf{h}', \mathbf{h}'')$, where prime indicates known disease genotype and double prime unknown disease genotype. Then $P(\mathbf{g}|\mathbf{x}) = P(\mathbf{g}|\mathbf{h}') = P(\mathbf{g})/(\mathbf{h}') = P(\mathbf{g})/\Sigma P(\mathbf{h})$, where the sum runs over all unobserved disease genotype vectors $\mathbf{h}''$. Using assumption (b) above and the law of total probability, the denominator may be evaluated as

$$\sum_{\mathbf{h}''} P(\mathbf{h}) = \prod_{k\epsilon O'} P(h_k) \prod_{l\epsilon D'} P(h_l|h_{l_1}, h_{l_2}) \sum_{\mathbf{h}''} \prod_{k\epsilon O''} P(h_k) \prod_{l\epsilon D''} P(h_l|h_{l_1}, h_{l_2})$$

$$= \prod_{k\epsilon O'} P(h_k) \prod_{l\epsilon D'} P(h_l|h_{l_1}, h_{l_2}) .$$

Finally, noting that $P(g_l|g_{l_1}, g_{l_2})/P(h_l|h_{l_1}, h_{l_2}) = P(g_l|g_{l_1}, g_{l_2}, h_l)$, and $P(g_k)/P(h_k) = P(g_k|h_k)$,

$$P(\mathbf{g}|\mathbf{x}) = \frac{\prod_{k\epsilon O'} P(g_k) \prod_{l\epsilon D'} P(g_l|g_{l_1}, g_{l_2}) \prod_{k\epsilon O''} P(g_k) \prod_{l\epsilon D''} P(g_l|g_{l_1}, g_{l_2})}{\sum_{\mathbf{h}''} P(\mathbf{h})}$$

$$= \prod_{k\epsilon O'} P(g_k|h_k) \prod_{l\epsilon D'} P(g_l|g_{l_1}, g_{l_2}, h_l) \prod_{k\epsilon O''} P(g_k) \prod_{l\epsilon D''} P(g_l|g_{l_1}, g_{l_2}).$$

Using assumption (a) and the fact that conditioning on unknown phenotype does not alter $P(g_k)$ or $P(g_l|g_{l_1}, g_{l_2})$, the result is proven.

APPENDIX B

STANDARD ERRORS

The estimators of $P[M_d^*(t) \geq c]$ in equation (3) are of the form $1 - \hat{p}_1^2 \hat{p}_2 (t = 0)$, $1 - \hat{p}_1\hat{p}_2$ $(0 < t < d/2)$, and $1 - \hat{p}_1^2$ $(t = d/2)$, where $\hat{p}_1$ and $\hat{p}_2$ are independent sample

proportions. Similarly, the estimator in equation (5) of the probability of at least one pedigree maximum lod score greater than c is of the form

$$1 - \prod_{k=1}^{K} \hat{p}_{1k}^2 \hat{p}_{2k} \; (t = 0), \; 1 - \prod_{k=1}^{K} \hat{p}_{1k} \hat{p}_{2k} \left(0 < t < \frac{d}{2}\right),$$

and

$$1 - \prod_{k=1}^{K} \hat{p}_{1k}^2 \left(t = \frac{d}{2}\right).$$

Note that the $\{\hat{p}_{ik}\}$ are independent random variables and that $N\hat{p}_{ik}$ is distributed as binomial on N trials with probability of success equal to the corresponding population proportion p_{ik}.

If Z is distributed as binomial on N trials with probability of success on each trial p, then

$$\begin{aligned} \mathrm{E}(Z) &= Np \\ \mathrm{E}(Z^2) &= N(N-1)p^2 + Np \\ \mathrm{E}(Z^4) &= g(N,p) = N(N-1)(N-2)(N-3)p^4 \\ &\quad + 6N(N-1)(N-2)p^3 + 7N(N-1)p^2 + Np \,. \end{aligned}$$

Hence, using the definition of variance and the independence of the $\{\hat{p}_{ik}\}$,

$$\mathrm{Var}\left[1 - \prod_{k=1}^{K} \hat{p}_{lk}^2 \hat{p}_{2k}\right] = N^{-5K} \prod_{k=1}^{K} \left\{ g(N, p_{1k}) \left[(N-1)p_{2k}^2 + p_{2k} \right] \right\} - N^{-2K} \prod_{k=1}^{K} \left\{ \left[(N-1)p_{1k}^2 + p_{1k} \right] p_{2k} \right\}^2 .$$

$$\mathrm{Var}\left[1 - \prod_{k=1}^{K} \hat{p}_{1k} \hat{p}_{2k}\right] = N^{-2K} \prod_{i=1}^{2} \prod_{k=1}^{K} \left[(N-1)p_{ik}^2 + p_{ik} \right] - \left[\prod_{i=1}^{2} \prod_{k=1}^{K} p_{ik} \right]^2 \qquad \text{(A-1)}$$

$$\mathrm{Var}\left[1 - \prod_{k=1}^{K} \hat{p}_{1k}^2\right] = N^{-4K} \prod_{k=1}^{K} g(N, p_{1k}) - N^{-2K} \left\{ \prod_{k=1}^{K} \left[(N-1)p_{1k}^2 + p_{1k} \right] \right\}^2 .$$

Substituting $\hat{p}_{ik}$ for p_{ik} in on the right-hand sides of equations (A-1) and taking square roots gives the desired standard errors.

REFERENCES

1. RENWICK JH: The mapping of human chromosomes. *Ann Rev Genet* 5:81–120, 1971
2. ELSTON RC, LANGE K: The prior probability of autosomal linkage. *Ann Hum Genet* 38:341–350, 1975
3. BOTSTEIN D, WHITE RL, SKOLNICK M, DAVIS RW: Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331, 1980
4. LANGE K, BOEHNKE M: How many polymorphic genes will it take to span the human genome? *Am J Hum Genet* 34:842–845, 1982
5. BISHOP DT, CANNINGS C, SKOLNICK M, WILLIAMSON JA: The number of polymorphic DNA clones required to map the human genome, in *Stastical Analysis of DNA Sequence Data,* edited by WEIR BS, New York, Marcel Dekker, 1983, pp 181–200
6. 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
7. OTT J: A computer program for linkage analysis of general human pedigrees. *Am J Hum Genet* 28:528–529, 1976
8. HASSTEDT SJ, CARTWRIGHT PE: PAP-*Pedigree Analysis Package*. Technical Report No. 13. Salt Lake City, Univ. of Utah, Department of Medical Biophysics and Computing, 1981
9. LATHROP GM, LALOUEL JM: Easy calculations of lod scores and genetic risks on small computers. *Am J Hum Genet* 36:460–465, 1984
10. DAHLQUIST G, BJORCK A, ANDERSON N: *Numerical Methods*. Englewood Cliffs, N.J., Prentice-Hall, 1974, pp 266–267
11. MORTON NE: Sequential tests for the detection of linkage. *Am J Hum Genet* 7:277–318, 1955
12. CHOTAI J: On the lod score method in linkage analysis. *Ann Hum Genet* 48:359–378, 1984
13. SKOLNICK MH, BISHOP DT, CANNINGS C, HASSTEDT SJ: The impact of RFLPs on human gene mapping, in *Genetic Epidemiology of Coronary Heart Disease: Past, Present, and Future,* edited by RAO DC, ELSTON RC, KULLER LH, FEINLEIB M, CARTER C, HAVLIK R, New York, Alan R. Liss, 1984, pp 271–292
14. THOMPSON EA, KRAVITZ K, HILL J, SKOLNICK MH: Linkage and the power of a pedigree structure, in *Genetic Epidemiology,* edited by MORTON NE, CHUNG CS, New York, Academic Press, 1978, pp 247–253
15. ELSTON RC, BONNEY GE: Sampling considerations in the design and analysis of family studies, in *Genetic Epidemiology of Coronary Heart Disease: Past, Present, and Future,* edited by RAO DC, ELSTON RC, KULLER LH, FEINLEIB M, CARTER C, HAVLIK R, New York, Alan R. Liss, 1984, pp 349–371
16. OTT J: *Analysis of Human Genetic Linkage*. Baltimore, Johns Hopkins Univ. Press, 1985, pp 40–56
17. RAO CR: *Linear Statistical Inference and Its Applications*. New York, John Wiley, 1973, pp 329–332