Genetic Epidemiology 13:117–137 (1996)

Affected-Sib-Pair Interval Mapping and Exclusion for Complex Genetic Traits: Sampling Considerations

Elizabeth R. Hauser, Michael Boehnke, Sun-Wei Guo, and Neil Risch

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor (E.R.H., M.B., S.-W.G.); Department of Genetics, School of Medicine, Stanford University, Stanford, California (N.R.)

We describe an extension of Risch's [(1990a,b) *Am J Hum Genet* 46:222–228, 229–241] method of linkage detection and exclusion for complex genetic traits. The method uses interval mapping to infer disease locus identity-by-descent (IBD) sharing for affected sib pairs (ASPs) based on marker information for the ASP and other genotyped family members. The method is likelihood based, and makes use of Risch's parameterization in terms of recurrence risk ratios for relatives. We describe specific linkage detection and exclusion tests for use as genome screening tools to prioritize genomic regions for further study. We also examine issues of optimal study design.

We advocate initially typing a large panel of ASPs (and no additional family members) with a map of genetic markers evenly spaced at 10–20-cM intervals. We recommend a screening procedure that 1) investigates further all regions with maximum lod scores greater than 1 and 2) excludes from consideration those regions that result in lod scores less than -2 at the smallest genetic effect that is viewed as important to detect. Further investigation of an interval might include typing other available families or family members, typing additional markers in the interval, and carrying out further statistical analyses. This strategy is efficient in the number of genotypings required and focuses attention on regions most likely to harbor a disease gene with a substantial impact on disease risk, while resulting in the pursuit of a manageable number of false-positive linkage results. Modification may be required if insufficient ASPs are available or if families come from a significantly admixed population.

Key words: genetic linkage, gene mapping, lod score, recurrence risk

Received for publication July 26, 1995; revision accepted October 23, 1995.

Address reprint requests to Michael Boehnke, Ph.D., Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029.

INTRODUCTION

There have been many recent successes in the mapping and positional cloning of human disease genes. However, most such successes have involved simple Mendelian diseases with clear modes of inheritance. In contrast, many diseases exhibit familial aggregation but fail to exhibit any simple Mendelian mode of inheritance. Such complex genetic diseases include most of the diseases that have a substantial impact on the public health, including heart disease, many forms of cancer, hypertension, diabetes, and schizophrenia. Identification of the genes involved in complex genetic diseases would have substantial impact on our understanding of disease etiology and on the prevention of disease occurrence. Because complex diseases likely are due to interplay of multiple genetic and environmental factors, mapping and cloning the genes for complex diseases is likely to be very difficult. However, recent successes such as the mapping of loci for insulin-dependent diabetes [Davies et al., 1994; Field et al., 1994; Hashimoto et al., 1994] and the identification of a role for apolipoprotein E in Alzheimer disease [Strittmatter et al., 1993] provide a basis for optimism.

The methods used in previous linkage studies fall into two general categories: mode-of-inheritance–based likelihood methods such as lod scores [Haldane and Smith, 1947; Morton, 1955] and location scores [Lathrop et al., 1984] and non-parametric mode-of inheritance–free methods such as sib pair [Penrose, 1935; Suarez et al., 1978] and affected-relative-pair [Weeks and Lange, 1988, 1992; Risch 1990b; Bishop and Williamson, 1990] methods.

Mode-of-inheritance–based methods have the advantages of 1) being statistically efficient when the mode of inheritance is known, 2) providing an estimate of disease gene location, and 3) permitting both linkage detection and exclusion. The major problem in applying these methods to complex diseases is the choice of mode of inheritance. Performing linkage analysis using several different modes of inheritance is one solution; however, this approach can result in increased probability of falsely concluding linkage, or alternatively in a loss of power, and often the selection of genetic models can be rather ad hoc.

Mode-of-inheritance–free methods are particularly suited to complex genetic diseases for which mode of inheritance generally is unknown. These methods use smaller subsets of family members, usually relative pairs, so that sampling family members generally is easier than when extended pedigrees are sought. Such methods are computationally simple and easy to apply. However, traditional mode-of-inheritance–free methods generally have not been used to provide an estimate of disease gene location or to perform exclusion mapping, although in principle they could [Risch, 1990b, 1993].

In this paper we extend Risch's mode-of-inheritance–free method of linkage analysis parameterized by the recurrence risk ratios for first-degree relatives of an affected individual. This method is intermediate between the traditional mode-of-inheritance–based lod score method and non-parametric, mode-of inheritance–free relative pair methods. Like the traditional non-parametric methods, it does not require specification of the (unknown) mode of inheritance. Like standard mode-of-inheritance–based methods, it provides an estimate of disease gene location and permits both linkage mapping and exclusion. We employ a maximum likelihood interval mapping approach to test for linkage detection and exclusion and to estimate model parameters. Previous methods assumed knowledge of identity-by-descent

(IBD) status as might be determined given marker data on the affected sib pair (ASP) and their parents. In the context of these methods we examine the information gained by including parents or sibs (when parents are unavailable) in addition to the ASP. We evaluate power and efficiency of different constellations of family members, map densities, and marker informativities and evaluate critical values for linkage detection and exclusion. Finally, we propose a general strategy for mapping genes for complex genetic traits.

We find that when a large number of ASPs are easy to obtain, an efficient study design for linkage detection is to type only the ASPs and no other family members for relatively widely spaced markers, say 10–20 cM. We recommend a genome screening procedure which investigates all intervals with lod scores greater than 1 by typing additional families and family members and additional markers in those intervals [see also Elston, 1992]. Our simulations suggest that such a procedure results in an acceptably low false positive rate and exclusion of the majority of the intervals in which a disease locus is not present. Modification of this strategy may be required if insufficient ASPs are available, if families come from a significantly admixed population, or if typing only the ASPs compromises accurate allele scoring. When the number of ASPs is limited, the information for linkage detection can be increased by typing additional family members of the ASPs.

MATERIALS AND METHODS

Model and Data

The genetic disease model we use was first described by Risch [1990a,b]. The model is parameterized by the recurrence risk ratio λ_R that compares disease risk in a relative of type R of an affected individual to the disease risk of an individual chosen at random from the population, that is the population prevalence. For example, R might represent a sibling, offspring, grandchild, or cousin. For these different relative classes, λ_R can be estimated in epidemiologic studies of the disease of interest. By parameterizing the model in terms of the recurrence risk ratio, specification of a more detailed mode-of-inheritance model can be avoided. Such an approach is intermediate between traditional relative pair techniques which include no genetic model parameters and standard likelihood-based methods that require full specification of the mode of inheritance.

For a complex genetic disease, multiple loci may be involved in disease etiology. Each locus may be sufficient to cause disease, or the loci may show more complex interactions. Risch [1990a] describes three such models of gene interaction. Here we assume that the marginal effect of at least one trait locus is detectable, separate from the other loci or environmental determinants with which it may interact, and let the parameter λ_R^* refer to the disease recurrence risk ratio to relative class R conferred by a *specific* susceptibility locus. For a single-locus Mendelian trait with no sporadic cases, $\lambda_R = \lambda_R^*$. In our investigations, we concentrate on the case when λ_R^* is of modest size, since for complex diseases this often will be the case. In the remainder of this paper we drop the distinction between λ_R and λ_R^*, with the understanding that in what follows the λ's refer to locus-specific recurrence risk ratios.

In this paper we consider data on ASPs and additional members of the nuclear families of the ASPs, and so concentrate on λ_s and λ_o, the recurrence risk ratios to

siblings and offspring of affected individuals, respectively. Let z_i be the probabilities that an ASP shares i genes identical by descent (IBD) at the disease locus. Risch [1987, 1990b] showed that for any disease locus the probabilities $\mathbf{z} = [z_0, z_1, z_2]$ satisfy $\mathbf{z} = [0.25/\lambda_s, 0.50\lambda_o/\lambda_s, 1 - 0.25/\lambda_s - 0.50\lambda_o/\lambda_s]$. Assuming a genetic model in which the alleles at the disease locus act additively, $\lambda_o = \lambda_s = \lambda$, and $\mathbf{z} = [0.25/\lambda, 0.50, 0.50 - 0.25/\lambda]$. When the specific locus is not related to the disease of interest, $\lambda_o = \lambda_s = 1$ and $\mathbf{z} = [1/4, 1/2, 1/4]$. By typing genetic markers throughout the genome in ASPs and perhaps in additional family members, we seek to identify intervals in which the ASPs demonstrate elevated IBD sharing. For this purpose, we employ a likelihood-based interval-mapping approach for linkage mapping and exclusion mapping of complex diseases. For ease of explication, in much of what follows we assume an additive model, a map of codominant markers with known distances between the markers, and no genetic interference, although these assumptions are not all necessary (see Discussion).

Test Statistics for Linkage Detection and Exclusion

Consider the case of a disease locus flanked by two genetic markers (Fig. 1). Here we use only intervals defined by pairs of flanking markers, rather than a full multipoint analysis (but see Discussion). Let θ be the recombination fraction between the flanking markers, and let θ_1 and θ_2 be the recombination fractions between the disease locus and the first and second flanking markers, respectively. Assuming no interference, $\theta_2 = (\theta - \theta_1)/(1 - 2\theta_1)$. Let $\mathbf{z} = [z_0, z_1, z_2]$ represent the true IBD sharing distribution at the disease locus in ASPs, and let $\mathbf{X}_n$ be the marker phenotypes for all genotyped members of family n. To detect or exclude linkage, we calculate $P(\mathbf{X}_n \mid \text{ASP}; \mathbf{z}, \theta_1, \theta)$, the probability of the marker data $\mathbf{X}_n$ on family n conditional on the disease status of the ASP, since it is this disease information that brought the family to our attention. We defer calculation of this probability to the next section.

Given N independent ASP families, we define the lod score as

$$\text{lod}(\mathbf{z}, \theta_1; \theta) = \sum_{n=1}^{N} \log_{10} \left(\frac{P(\mathbf{X}_n \mid \text{ASP}; \mathbf{z}, \theta_1, \theta)}{P(\mathbf{X}_n \mid \text{ASP}; \mathbf{z} = [1/4, 1/2, 1/4], \theta_1, \theta)} \right).$$

This lod score compares the likelihood of the data when there is a disease locus at a given location to the likelihood of the data when there is no disease locus in the interval [see Risch, 1990b]; it provides the basis to detect or to exclude linkage.

To test for linkage in a particular interval we maximize the lod score over $\mathbf{z}$ and θ_1. We consider θ_1 values in the range $0 \leq \theta_1 \leq \theta$ and $\mathbf{z}$ values in the "possible

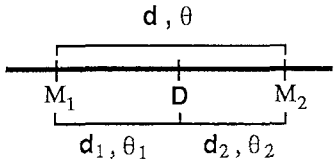

Fig. 1. Definitions of interval mapping parameters: M_1, M_2, genetic markers flanking the disease locus D; θ, θ_1, and θ_2, recombination fractions between the flanking markers and between the disease locus and marker loci; and d, d_1, and d_2, corresponding map distances.

triangle" representing models which are consistent with the effects of a single genetic locus on disease risk [Holmans, 1993]. This triangle is bounded by the lines $z_0 = 0$, $z_1 = 1/2$, and $z_2 = 2z_0$. For an additive model $z_1 = 1/2$, and maximization is restricted to $0 \leq z_0 \leq 1/4$ (or equivalently $1/4 \leq z_2 \leq 1/2$). For the additive model, we may parameterize in terms of $\lambda = 0.25/z_0$ and replace $\mathbf{z} = [1/4, 1/2, 1/4]$ by $\lambda = 1.0$. Large, positive lod scores suggest the presence of a disease locus.

For exclusion mapping, we calculate the lod score as a function of θ_1 and $\mathbf{z}$, or under additivity, θ_1 and λ. At a specific value of λ, say λ_E, we exclude those parts of the interval which result in sufficiently negative lod scores and build an exclusion map of the genome, indicating those regions unlikely to contain a locus with marginal effect λ_E or greater [Risch, 1993]. Alternatively, we could fix θ_1 or the portion of the interval of interest and then determine the λ values which might be excluded. The same approaches can be taken for the more general model parameterized by $\mathbf{z}$.

At the null hypothesis value $\mathbf{z} = [1/4, 1/2, 1/4]$, the parameter space is degenerate so that the recombination fraction between the disease locus and the flanking marker is not identifiable. As a result, when $\mathbf{z} = [1/4, 1/2, 1/4]$ the asymptotic distribution of the likelihood ratio statistic is unknown (see Discussion). We use simulation to examine the empirical distribution of the maximum lod score.

Likelihood

To calculate the probability $\mathrm{P}(\mathbf{X} \mid \mathrm{ASP}; \mathbf{z}, \theta_1, \theta)$ of the marker data $\mathbf{X}$ for a family conditional on the disease status of the ASP, we condition on IBD sharing by the ASP at the disease and marker loci. Let i_{kj} be 1 if the ASP shares an allele IBD from parent k = m(other) or f(ather) at locus j, and 0 otherwise. Here, $j = 1$, 2, or D for the first marker locus, second marker locus, and disease locus, respectively, Then $I_D = (i_{mD}, i_{fD})$ and $I_M = (i_{m1}, i_{m2}, i_{f1}, i_{f2})$ completely describe the IBD sharing at the disease and marker loci, respectively. As an example, consider the families in Figure 2. In family 1 both parents are genotyped. At the first marker, the ASP share the father's allele IBD but not the mother's allele. At the second marker, both parents are homozygous and the IBD allele sharing status in the ASP is unknown. Thus $I_M \in \{(0,0,1,0), (0,1,1,0), (0,0,1,1), (0,1,1,1)\}$. In family 2, if the unaffected sibling is not included, all 16 4-tuples indicating ASP IBD sharing are possible. Adding the genotype of the unaffected sibling allows the parental marker genotypes and ASP IBD status to be inferred unambiguously and $I_M = (1,1,1,1)$.

Conditioning on the IBD status of the ASP we can write the likelihood as

$$\begin{aligned}\mathrm{P}(\mathbf{X} \mid \mathrm{ASP}; \mathbf{z}, \theta_1, \theta) &= \sum_{I_M} \mathrm{P}(\mathbf{X} \mid I_M; \theta) \sum_{I_D} \mathrm{P}(I_D \mid \mathrm{ASP}; \mathbf{z})\, \mathrm{P}(I_M \mid I_D; \theta_1, \theta) \\ &= \sum_{I_M} \mathrm{P}(\mathbf{X}, I_M; \theta) H(I_M; \mathbf{z}, \theta_1, \theta)\end{aligned}$$

where

$$H(I_M; \mathbf{z}, \theta_1, \theta) = \sum_{I_D} \mathrm{P}(I_D \mid \mathrm{ASP}; \mathbf{z})\, \mathrm{P}(I_M \mid I_D; \theta_1, \theta) / \mathrm{P}(I_M; \theta)$$

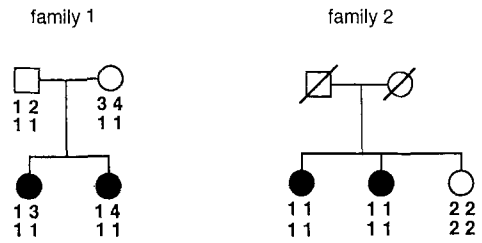

Fig. 2. Example of affected-sib-pair nuclear families.

[Hauser et al., 1994]. An analogous expression was obtained independently by Olson [1995a].

The terms in the expression $H(I_M; \mathbf{z}, \theta_1, \theta)$ are simple functions of the parameters. By definition

$$P(I_D \mid \text{ASP}; \mathbf{z}) = \begin{cases} z_0 & \text{if } I_D = (0,0) \\ z_1/2 & \text{if } I_D = (0,1),(1,0). \\ z_2 & \text{if } I_D = (1,1) \end{cases}$$

Under the assumptions of no inbreeding and no interference

$$P(I_M \mid I_D; \theta_1, \theta) = \prod_{k=m,f} \prod_{j=1}^{2} P(i_{kj} \mid i_{kD})$$

where

$$P(i_{kj} \mid i_{kD}) = \begin{cases} \theta_j^2 + (1-\theta_j)^2 & \text{if } i_{kj} = i_{kD} \\ 2\theta_j(1-\theta_j) & \text{if } i_{kj} \neq i_{kD} \end{cases} \quad \text{and} \quad \theta_2 = \frac{\theta - \theta_1}{1 - 2\theta_1}.$$

Similarly, $P(I_M; \theta) = \prod_{k=m,f} P(i_{k1}) P(i_{k2} \mid i_{k1})$, where $P(i_{k1}) = 1/2$ and

$$P(i_{k2} \mid i_{k1}) = \begin{cases} \theta^2 + (1-\theta)^2 & \text{if } i_{k1} = i_{k2} \\ 2\theta(1-\theta) & \text{if } i_{k1} \neq i_{k2} \end{cases}.$$

H incorporates all of the information about the disease model and the position of the disease locus relative to the markers. In addition, H is trivial to calculate, and only $2^4 = 16$ terms are needed for each combination of z and θ_1.

Calculation of $P(\mathbf{X}, I_M; \theta)$ is somewhat more complicated, but it is readily accomplished by using a variation of the Elston-Stewart [1971] algorithm; further, it needs to be carried out only once per family, independent of the number of $\mathbf{z}$ and θ_1 values to be considered. We restrict our attention to the case of nuclear families, although generalization to extended pedigrees is possible. Let $\mathbf{g} = (g_m, g_f, g_1, g_2, \ldots, g_s)$ be a two-marker genotype vector for a family where m

and f are the mother and father, 1 and 2 are the ASP, and 3 through s are additional siblings. Then

$$\begin{aligned} &\mathrm{P}(\mathbf{X}, I_M; \theta) \\ &\quad = \sum_{\mathbf{g}} \mathrm{P}(\mathbf{g}; \theta)\, \mathrm{P}(X \mid \mathbf{g})\, \mathrm{P}(I_M \mid \mathbf{g}; \theta) \\ &\quad = \sum_{g_m, g_f} \mathrm{P}(g_m)\, \mathrm{P}(g_f) Q(g_m, g_f) \sum_{g_1, g_2} \prod_{a=1}^{2} \mathrm{P}(g_a \mid g_m, g_f) \prod_{k=m,f} \mathrm{P}(i_{k1}, i_{k2} \mid g_1, g_2, g_k) \quad (1) \end{aligned}$$

where

$$Q(g_m, g_f) = \prod_{n=3}^{s} \sum_{g_n} \mathrm{P}(g_n \mid g_m, g_f)$$

is the function that results from peeling the s-2 non-ASP siblings onto the parents. $\mathrm{P}(g_k)$ is the prior probability of genotype g_k, and $\mathrm{P}(g_i \mid g_m, g_f)$ is the (transmission) probability of offspring genotype g_i given parental genotypes g_m and g_f. Missing from (1) are the penetrance terms $\mathrm{P}(\mathbf{X} \mid \mathbf{g})$. As described below, we carry out the summations only for those genotypes consistent with the observed marker phenotypes, so that the penetrance terms are all one and hence can be ignored.

The last term in (1) is the product of the conditional marker IBD probabilities given the genotypes of the ASP and a parent. Table I specifies these simple probabilities as a function of θ and of parent genotypes and the identity-by-state (IBS) relationship of the ASP alleles for each marker. Returning to example family 1 (Fig. 2), the father's and mother's contributions are described by lines 9 and 11 of the table, respectively. For family 2, ignoring the genotypes of the unaffected sibling, the marker genotypes of both parents are unknown; all parental genotypes g_k consistent with the genotypes of the ASP must be considered, in this case lines 1, 5, 9, and 13. Taking into account the unaffected sibling reduces the number of possible parental genotypes to the case represented by line 13.

To reduce computation, the parental genotype sum in (1) can be taken only over parental marker genotypes compatible with observed marker phenotypes; similarly the offspring sum can be taken over the intersection of the set of parental genotypes compatible with the set of observed offspring marker phenotypes. This strategy can substantially reduce the number of genotypes to be considered. At best, when parents and the ASP are genotyped, the set of compatible parental genotypes may include no more than four ordered two-marker-locus genotypes. At worst, when only the ASP is genotyped, given appropriate allele recoding, there are no more than $9^2 = 81$ possible ordered two-marker-locus genotypes and $81^2 = 6561$ two-marker-locus mating types.

Simulations

We carried out computer simulations to examine power and size for linkage detection and exclusion, and the choice of critical values which might be used for a genome scan. For ease of explication, we focus on additive models for which

TABLE I. Conditional Identity-by-Descent (IBD) Probabilities for an Affected Sib Pair (ASP) at the Flanking Marker Loci Given Genotypes for One Parent and the ASP

Parental genotype g_k homozygous at		ASP share parent k allele IBS at		Conditional marker IBD probabilities[a]			
Marker 1?	Marker 2?	Marker 1?	Marker 2?	P(0, 0)	P(0, 1)	P(1, 0)	P(1, 1)
Yes	Yes	Yes	Yes	$\psi/2$	$(1-\psi)/2$	$(1-\psi)/2$	$\psi/2$
			No	—	—	—	—
		No	Yes	—	—	—	—
			No	—	—	—	—
	No	Yes	Yes	0	$1-\psi$	0	ψ
			No	ψ	0	$1-\psi$	0
		No	Yes	—	—	—	—
			No	—	—	—	—
No	Yes	Yes	Yes	0	0	$1-\psi$	ψ
			No	—	—	—	—
		No	Yes	ψ	$1-\psi$	0	0
			No	—	—	—	—
	No	Yes	Yes	0	0	0	1
			No	0	0	1	0
		No	Yes	0	1	0	0
			No	1	0	0	0

[a] $P(a,b) = P(i_{k1} = a, i_{k2} = b \mid g_1, g_2, g_k)$ where i_{kj} is the indicator of ASP IBD allele sharing for parent k at marker j and g_1, g_2, and g_k are the two-locus genotypes for the ASP and parent k; $\psi = \theta^2 + (1-\theta)^2$; —, impossible case.

$\lambda_o = \lambda_s = \lambda$ and $\mathbf{z} = [0.25/\lambda, 0.50, 0.50 - 0.25/\lambda]$. Since the locus-specific recurrence risk ratios λ for a complex disease often will be low, we consider λs of 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, and 3.0. For $N = 100$, 200, 400, or 800 ASP families, we consider the information provided by genotyping the ASP together with $N_p = 0$, 1, or 2 parents and $N_s = 0$ to 6 additional sibs. We also consider the impact of distance between flanking markers ($d = 2$, 5, 10, 20 cM), marker informativity (two, four, or ten equally frequent alleles, heterozygosity = 0.50, 0.75, or 0.90), and disease locus position (in the middle of the interval and near a flanking marker).

It has been noted by several investigators that errors in specification of allele frequencies and particularly population admixture can be serious problems in linkage studies, especially in affected-relative-pair studies [Weeks and Lange, 1992; Risch, 1992; Freimer et al., 1993; Holmans, 1993]. To address this issue, we simulated a sample of 400 ASPs from a mixture of two populations with different marker allele frequencies. For this simulation we used markers with eight alleles and allele frequencies 0.36, 0.13, 0.13, 0.13, 0.08, 0.08, 0.08, and 0.01 (heterozygosity = 0.80) in the first population, and with the frequencies either for the first two alleles or for the first and fifth alleles switched in the second population. A fraction α of the ASP families were chosen from the first population, the remaining fraction $1 - \alpha$ from the second. To analyze these admixed data we assumed the averaged allele frequencies from the combined population.

To simulate marker genotypes for the family members, we simulated genotypes first for the parents, next for the ASP conditional on the genotypes of the parents

and the parameters θ_1 and θ and λ, and finally for the additional siblings conditional on the genotypes of the parents and on θ. To facilitate comparisons of linkage information provided by different family structures, we generated data for ten-person nuclear families but used marker genotypes from the subset of individuals of interest. This scheme allowed a more efficient comparison of different sampling strategies by reducing sampling variation.

For each simulation condition we generated 1,000 replicate data sets. For each data set we calculated the maximum lod score and the maximum likelihood estimates of θ_1 and λ. In addition, for each of several values of λ, we calculated the proportion of the interval excluded and noted whether the entire interval was excluded.

RESULTS

We begin by examining the expected maximum lod score and parameter estimates for different sample sizes, family structures, map densities, and marker heterozygosities. Next we present size and power of tests for linkage detection and linkage exclusion based on the empirical distribution of the lod score. Finally we examine the impact of errors in the marker map on linkage detection.

Effect of Typing Additional Family Members on the Maximum Lod Score

Table II shows the expected maximum lod score for several values of the recurrence risk ratio for first degree relatives λ. Not surprisingly, as λ increases, the expected maximum lod score increases so that disease loci with larger recurrence risk ratios are more easily detected than loci with smaller recurrence risk ratios. For a fixed number of families N, the most information for linkage is obtained when both parents are genotyped ($N_p = 2$), the least when no parents are typed ($N_p = 0$). How-

TABLE II. Expected Maximum Lod Score as a Function of the Number of Families N, the Number of Typed Parents N_p, and the Recurrence Risk Ratio λ Given Markers With Four Equally-Frequent Alleles, a 10-cM Intermarker Distance, and Disease Locus in the Middle of the Interval*

		λ						
N	N_p	1.0	1.2	1.4	1.6	1.8	2.0	3.0
100	0	0.15	0.43	0.78	1.15	1.49	1.80	2.97
	1	0.16	0.47	0.84	1.26	1.65	1.99	3.31
	2	0.16	0.52	0.96	1.45	1.91	2.32	3.92
200	0	0.15	0.61	1.26	1.98	2.65	3.25	5.57
	1	0.14	0.65	1.37	2.18	2.93	3.60	6.22
	2	0.15	0.75	1.58	2.52	3.42	4.24	7.44
400	0	0.15	0.96	2.20	3.60	4.92	6.12	10.75
	1	0.15	1.05	2.42	4.01	5.49	6.84	12.10
	2	0.16	1.20	2.82	4.68	6.44	8.10	14.53
800	0	0.13	1.59	4.01	6.80	9.38	11.83	21.18
	1	0.14	1.76	4.46	7.62	10.54	13.30	23.92
	2	0.15	2.04	5.22	8.97	12.46	15.82	28.75

*Standard errors $\leq .0056\sqrt{N}$.

ever, on a per genotype basis, typing only the ASP is the most efficient strategy [see also Risch, 1992; Holmans, 1993]. The same number of genotypes are required when typing both parents and the ASP in N families as when typing just the ASP in $2N$ families, yet for the cases we considered, the expected maximum lod score was 17% to 46% larger when typing the $2N$ ASPs. Figure 3 shows the effect of typing siblings in addition to the ASP on the expected lod score for $N = 400$ nuclear families. Each additional sibling increases the expected maximum lod score, and typing six siblings in addition to the ASP provides almost as much information as typing both parents, but it is always better to type parents rather than siblings if the parents are available. Again, on a per genotype basis, it is more efficient to type more ASPs rather than to type additional siblings. Since typing parents and/or additional siblings was seen to be less efficient than typing only ASPs, in subsequent analyses we restrict our attention to only ASPs.

Effect of Intermarker Distance on the Maximum Lod Score

Table III shows the expected maximum lod score when typing $N = 100$ or 400 ASPs and no other family members. As expected, reducing the intermarker distance increases the expected maximum lod score; this increase is larger when the disease locus is in the middle of the interval than when it is near one of the markers. For example, for $N = 400$ ASPs and $\lambda = 2.0$, the expected maximum lod score increases

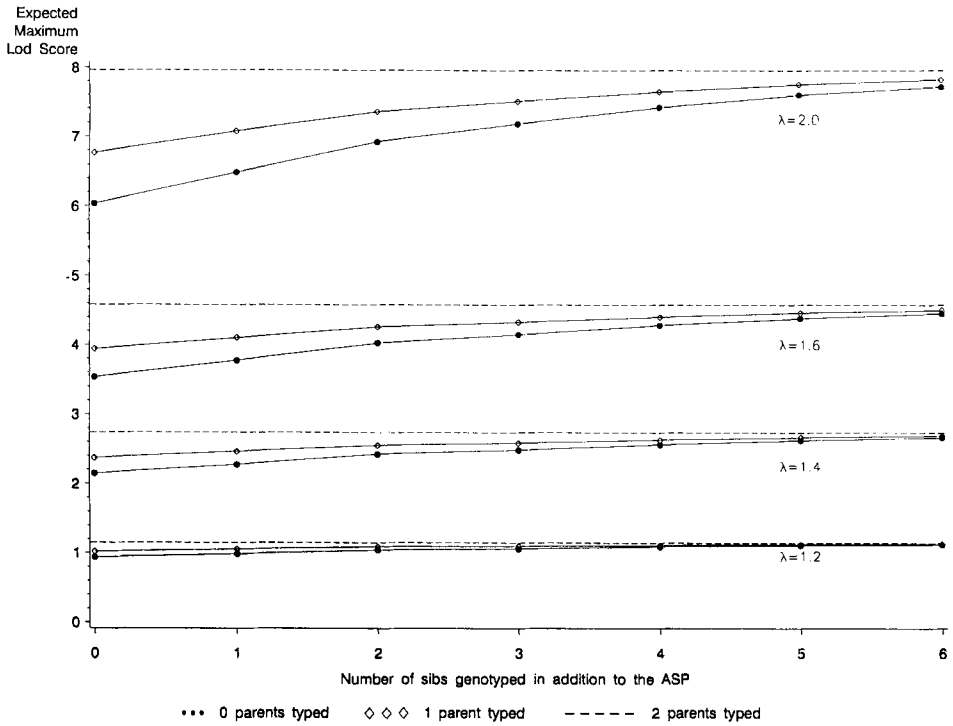

Fig. 3. The increase in the expected maximum lod score given genotypes on additional siblings for $N = 400$ families typed with markers with four equally frequent alleles, $d = 10$ cM, and disease locus in the middle of the interval.

TABLE III. Expected Maximum Lod Score and Means of the Maximum Likelihood Estimates of d_1, the Distance Between the First Marker and the Disease Locus and $z_0 = 0.25/\lambda$, the Probability That the ASP Shares Zero Disease Genes IBD, When Typing N ASPs With Markers With Four Equally Frequent Alleles*

				$N = 100$			$N = 400$		
d (cM)	λ	z_0	d_1	Lod score	$\hat{z}_0$	$\hat{d}_1$	Lod score	$\hat{z}_0$	$\hat{d}_1$
20	1.0	0.250	—	0.17	0.220	—	0.17	0.236	—
	1.4	0.179	10.0	0.67	0.171	9.4	1.70	0.179	10.0
			0.1	0.78	0.165	5.1	2.19	0.172	3.3
	2.0	0.125	10.0	1.40	0.124	9.8	4.47	0.126	10.0
			0.1	1.79	0.112	3.5	6.21	0.117	1.9
10	1.0	0.250	—	0.15	0.226	—	0.14	0.238	—
	1.4	0.179	5.0	0.78	0.172	4.5	2.20	0.177	4.8
			0.1	0.80	0.171	3.1	2.36	0.176	2.2
	2.0	0.125	5.0	1.80	0.122	4.8	6.12	0.125	4.9
			0.1	1.91	0.119	2.4	6.74	0.121	1.3
5	1.0	0.250	—	0.13	0.230	—	0.13	0.240	—
	1.4	0.179	2.5	0.82	0.174	2.4	2.53	0.177	2.4
			0.1	0.83	0.174	1.9	2.56	0.177	1.6
	2.0	0.125	2.5	2.00	0.123	2.3	7.22	0.124	2.4
			0.1	2.03	0.122	1.6	7.36	0.123	1.0
2	1.0	0.250	—	0.12	0.232	—	0.12	0.241	—
	1.4	0.179	1.0	0.86	0.176	0.9	2.73	0.178	1.0
			0.1	0.86	0.176	0.9	2.74	0.178	0.8
	2.0	0.125	1.0	2.15	0.124	1.0	8.00	0.124	1.0
			0.1	2.16	0.124	0.8	8.01	0.124	0.7

*Standard errors for the lod score $\leq .040$ for $N = 100$ and $\leq .081$ for $N = 400$. Standard errors for $\hat{z}_0 \leq 0.002$ for $N = 100$ and ≤ 0.001 for $N = 400$. Standard errors for $\hat{d}_1 \leq 0.30$ for $N = 100$ and ≤ 0.15 for $N = 400$.

from 4.47 when the interval size $d = 20$ cM to 8.00 for $d = 2$ cM. However, this less than twofold increase in linkage information comes at the expense of a tenfold increase in genotyping effort. Thus, given sufficient ASPs, a plausible genome screening strategy is to type markers at a relatively large intermarker distance, say $d = 10$ or 20 cM, and then to type more markers in intervals with interesting maximum lod scores (see below).

When the disease locus is close to one of the flanking markers, the expected maximum lod score is larger than when the disease locus is in the middle of the interval. For example, when $N = 400$, $\lambda = 2.0$, and $d = 20$ cM, the expected maximum lod score is 4.47 when the disease locus is in the middle of the interval and 6.21 when the disease locus is near a flanking marker. This difference is larger for larger intervals since there is more opportunity for recombination between the disease locus and both markers when the disease locus is in the middle.

Effect of Marker Informativity on the Maximum Lod Score

Table IV presents the expected maximum lod score for markers with two, four, or ten equally frequent alleles (heterozygosities 0.50, 0.75, or 0.90, respectively) given a

TABLE IV. Effect of Marker Informativity on the Expected Maximum Lod Score for Markers With Two, Four, or Ten Equally Frequent Alleles Given N = 400 ASPs, λ = 2.0, and d = 10 cM*

Number of alleles at		d_1 (cM)	
Marker 1	Marker 2	5.0	0.1
2	2	3.59	3.82
	4	5.08	4.71
	10	6.65	5.60
4	2	5.07	6.27
	4	6.12	6.74
	10	7.28	7.22
10	2	6.62	8.82
	4	7.27	8.98
	10	8.05	9.14

*Standard errors $\leq$ .087.

10-cM map. As expected, as marker heterozygosity increases, the expected maximum lod score increases. This increase is the most striking when the disease locus is near a flanking marker. For example, when the number of alleles at the flanking marker closest to the disease locus increases from 4 to 10, the expected maximum lod score increases from 6.74 to 8.98. The increase in linkage information due to an increase in marker heterozygosity can be observed even at a distance. When the number of alleles at the marker further from the disease locus is increased from 4 to 10, the expected maximum lod score increases from 6.74 to 7.22. Linkage detection is clearly enhanced by choosing the most polymorphic markers. Even so, markers that are only modestly polymorphic still can provide substantial linkage information.

Parameter Estimation

Table III shows parameter estimates $\hat{z}_0$ and $\hat{d}_1$ for several intermarker distances d, recurrence risk ratios $\lambda = 0.25/z_0$, and disease locus positions d_1. In general, the parameters are reasonably well estimated. For $\hat{z}_0$, the bias is generally small and decreases with increasing number of ASPs N. The bias in $\hat{z}_0$ is the greatest when $z_0 = 0.25$ ($\lambda = 1.0$) since the estimation procedure forces $z_0 \leq 0.25$ ($\lambda \geq 1.0$); even in this case the bias is generally small, particularly when d is small and N is large.

When the disease locus is in the middle of the interval, the bias in $\hat{d}_1$ is small and decreases as the interval size decreases. When the disease locus is near a flanking marker, the estimate of the disease locus position d_1 is biased toward the center of the interval. Again, this is to be expected since by using information on flanking markers only, we do not allow the disease locus to be outside the interval. Results for more or less polymorphic markers are similar (data not shown).

Size and Power to Detect Linkage

Table V shows the empirical size and power of the linkage detection test when using a lod score of 1, 2, or 3 as the critical value for declaring suggestive evidence of linkage. Even at a relatively low lod score critical value of $k = 1$, the probability

TABLE V. Probability of a Maximum Lod Score Greater Than k for N ASPs for Markers With Four Equally Frequent Alleles, a 10-cM Intermarker Distance, and Disease Locus in the Middle of the Interval

		λ						
N	k	1.0	1.2	1.4	1.6	1.8	2.0	3.0
100	1	0.024	0.13	0.30	0.48	0.61	0.71	0.93
	2	0.002	0.03	0.08	0.16	0.26	0.38	0.70
	3	0.000	0.00	0.02	0.05	0.09	0.15	0.45
200	1	0.022	0.21	0.53	0.78	0.90	0.95	1.00
	2	0.003	0.05	0.20	0.44	0.63	0.77	0.98
	3	0.000	0.01	0.06	0.19	0.36	0.51	0.90
400	1	0.025	0.38	0.81	0.97	1.00	1.00	1.00
	2	0.003	0.13	0.49	0.81	0.96	0.99	1.00
	3	0.000	0.03	0.23	0.59	0.82	0.94	1.00
800	1	0.020	0.65	0.98	1.00	1.00	1.00	1.00
	2	0.002	0.30	0.89	1.00	1.00	1.00	1.00
	3	0.000	0.12	0.69	0.97	1.00	1.00	1.00

of declaring linkage when there is no disease locus in the interval ($\lambda = 1.0$) is no larger than 0.025 for the cases considered. This represents no more than eight false-positives, on average, in a genome scan of 300 markers. A more stringent lod score critical value, say $k = 2$, results in a false-positive rate of no more than .003 or an average of one false-positive in a genome scan of 300 markers, but at the expense of a considerable loss of power for detecting linkage when a locus lies in the interval. Using $k = 1$, a sample size of $N = 400$ ASPs provides 81% power to detect a disease locus with recurrence risk ratio $\lambda = 1.4$; power is reduced to 49% or 23% for a critical value of 2 or 3. As always, the power to detect an effect of a given size is dependent on the sample size. A sample of $N = 100$ ASPs is sufficient to detect linkage of a locus with recurrence risk ratios $\lambda \geq 3.0$ with reasonable power for $k = 1$ or 2. $N = 200$–400 ASPs provide substantial power to detect a locus with $\lambda \geq 1.6$ for $k = 1$, and with $N = 800$ ASPs we may be able to detect a locus with λ as small as 1.2.

These results suggest that a reasonable strategy for a genome scan to localize a gene predisposing to a complex disease may be to screen the genome using a relatively low lod score critical value k, even as low as $k = 1$. A critical value of $k = 1$ gives excellent power at $\lambda \geq 1.4$ for $N \geq 400$ ASPs without resulting in a prohibitively large number of false-positives. While maximum lod scores as low as 1 certainly should not be regarded as conclusive evidence for linkage, they can indicate intervals most likely to harbor disease loci and worthy of further investigation [see also Elston, 1992].

Exclusion Mapping

An advantage of our parametric ASP linkage method is that it allows exclusion of intervals for which there is substantial evidence against the presence of a disease locus conferring a specified recurrence risk ratio λ_E [Risch, 1993]. A reasonable strategy is to exclude those locations with $\text{lod}(\lambda_E, \theta_1; \theta) \leq k$, for some $k < 0$. Table VI

TABLE VI. Percentage of Intervals Excluding the True Disease Gene With Recurrence Risk Ratio λ at Exclusion Recurrence Risk Ratios λ_E Using Lod ≤ k as Exclusion Criterion*

		λ_E							
		1.4		1.6		2.0		3.0	
λ	d (cM)	$k = -1$	$k = -2$	$k = -1$	$k = -2$	$k = -1$	$k = -2$	$k = -1$	$k = -2$
1.6	20	0.0	0.0	1.2	0.1	5.6	1.8	25.7	14.9
	10	0.0	0.0	0.3	0.1	3.7	1.2	25.2	18.1
	5	0.0	0.0	0.1	0.1	3.5	0.9	27.7	19.4
	2	0.0	0.0	0.2	0.0	3.5	1.2	29.5	21.4
2.0	20	0.0	0.0	0.0	0.0	0.4	0.1	3.8	1.2
	10	0.0	0.0	0.0	0.0	0.3	0.0	1.7	0.6
	5	0.0	0.0	0.0	0.0	0.1	0.0	1.6	0.5
	2	0.0	0.0	0.0	0.0	0.1	0.1	1.6	0.7

*N = 400 ASPs typed for markers with four equally-frequent alleles in a d cM intermarker distance, and disease locus in the middle of the interval.

displays the probabilities of lod scores less than $k = -1$ or -2 at the location of a disease gene for which $\lambda = 1.6$ or 2.0 when typing ASPs only. The percentages of falsely excluded intervals containing a disease locus with recurrence risk ratios $\lambda = 1.6$ or 2.0 is comfortingly small for $\lambda_E \leq \lambda$, even for the less stringent critical value $k = -1$. This suggests that this method will only rarely exclude an interval when that interval actually includes a disease locus with $\lambda \geq \lambda_E$. However, since excluding the disease locus is a very serious error, it may be preferable to use the more stringent critical value $k = -2$.

Table VII shows the results for exclusion mapping when the interval does not contain a disease locus ($\lambda = 1.0$) when we used as our exclusion criterion a lod score less than -2. As expected, as λ_E or N increases or d decreases, the percentage of each interval excluded and the probability that the entire interval is excluded both increase. For example, when typing $N = 400$ ASPs with an interval size of $d = 20$ cM, 85% of the intervals were fully excluded for $\lambda_E = 2.0$, while less than 60% were fully excluded for $\lambda_E = 1.6$. For $\lambda_E = 1.6$, the proportion of intervals fully excluded increased to 90% when $d = 2$ cM, suggesting that regions of the genome not completely excluded in an initial genome scan may be excluded when more markers are typed in the region.

Effect of Errors in Marker Map Parameters

Population admixture can cause serious problems for linkage analysis methods in which identity by descent must be inferred from identity by state. To address this issue, we simulated markers with eight alleles of unequal frequencies (Table VIII) and constructed samples in which a fraction α of the ASPs came from a population with one set of allele frequencies and the remaining fraction $1 - \alpha$ came from a population in which the frequencies of the most common allele and a second allele had been interchanged.

For the case in which no disease locus was present ($\lambda = 1.0$), admixture resulted in a shift of the maximum lod score distribution toward larger values. Thus, for any critical value k, the false-positive rate was substantially increased. For example, the

TABLE VII. Average Percentage of the Interval Excluded (% Ex) and Percentage of Replicates for Which the Entire Interval Is Excluded (% All) When No Disease Locus Is Present ($\lambda = 1$) Given N ASPs, a d cM Intermarker Distance, and Markers With Four Equally Frequent Alleles

		λ_E							
		1.4		1.6		2.0		3.0	
N	d	% Ex	% All	% Ex	% All	% Ex	% All	% Ex	% All
100	20	3.1	0.1	5.3	2.8	20.8	12.6	55.2	44.5
	10	1.9	1.5	10.9	7.9	33.6	27.6	69.9	64.0
	5	2.7	2.4	16.4	14.3	42.0	38.3	77.0	73.7
	2	3.6	3.3	20.6	19.3	49.4	47.5	82.1	80.6
200	20	7.0	3.6	29.4	20.7	58.8	48.8	87.7	81.4
	10	13.8	10.7	43.9	38.1	73.1	68.4	94.9	93.1
	5	18.3	16.1	52.4	49.0	80.7	78.5	96.4	95.5
	2	23.9	22.5	59.4	57.5	84.6	83.6	98.3	97.9
400	20	32.9	24.2	69.0	59.4	90.3	85.4	98.9	97.8
	10	49.0	43.9	82.1	78.5	95.7	94.5	99.8	99.7
	5	56.3	53.1	87.1	85.1	97.2	96.7	99.9	99.9
	2	63.7	62.0	90.8	90.3	98.0	97.7	99.9	99.9
800	20	70.8	61.4	93.9	90.7	99.2	98.5	100.0	100.0
	10	83.2	80.0	98.1	97.1	99.9	99.9	100.0	100.0
	5	88.3	86.6	98.9	98.7	99.9	99.9	100.0	100.0
	2	91.2	90.2	99.2	99.1	100.0	99.9	100.0	100.0

false-positive rate increased from 0.026 without admixture to 0.132 given $\alpha = 20\%$ admixture and allele frequency set B (see Table VIII) in the second population when we used critical value $k = 1$.

This increase was initially rather alarming, since for a 300-marker genome scan it would suggest the need to follow up approximately 40 false-positives. However, when a disease locus is present ($\lambda > 1.0$), admixture again results in a shift of the

TABLE VIII. Effect of Population Admixture α on the Probability of a Maximum Lod Score Greater Than k When There Is No Disease Locus*

			$\alpha = 0.20$		$\alpha = 0.50$	
λ	k	$\alpha = 0.00$	A	B	A	B
1.0	1	.026	.075	.132	.139	.257
	2	.002	.009	.019	.025	.063
	3	.000	.002	.004	.005	.012
1.4	1	.874	.950	.972	.977	.989
	2	.596	.795	.879	.871	.944
	3	.314	.534	.676	.664	.824
2.0	1	1.000	1.000	1.000	1.000	1.000
	2	.995	.998	1.000	1.000	1.000
	3	.976	.995	.999	.997	1.000

*$N = 400$, $d = 10$ cM distance between markers with eight alleles with frequencies 0.36, 0.13, 0.13, 0.13, 0.08, 0.08, 0.08, 0.01 and two sets of allele frequencies in an $(\alpha : 1 - \alpha)$ admixed population. A: Allele frequencies in the second population are 0.13, 0.36, 0.13, 0.13, 0.08, 0.08, 0.08, 0.01, respectively. B: Allele frequencies in the second population are 0.08, 0.13, 0.13, 0.13, 0.36, 0.08, 0.08, 0.01, respectively.

maximum lod score distribution toward larger values and seems to do so in a way entirely parallel to that for the no linkage case. For example, $k = 1$ for the no admixture case and $k = 2$ for the case of $\alpha = 50\%$ admixture and allele frequency set A result in similar false-positive rates (.026 and .025, respectively). Using these same critical values also results in essentially identical power to detect linkage: .874 and .871 when $\lambda = 1.4$, and 1.000 and 1.000 when $\lambda = 2.0$. Thus, by appropriate choice of the critical value k, it appears to be possible to achieve the same false-positive rate without reducing power to detect linkage (see Discussion).

Throughout we have assumed that the intermarker distance d is known without error. To test the effect of inaccurately specified d on linkage detection, we simulated data for $d = 10$ cM and then analyzed the data assuming various inflated or deflated values of d. We assumed a locus with risk $\lambda = 2.0$ ($z_0 = 0.125$), $N = 400$ ASPs, four equally frequent marker alleles, and disease locus in the middle of the interval. When d was assumed to be 5 cM, the expected maximum lod score decreased from 6.12 to 4.86 and the average $\hat{z}_0$ increased from 0.129 to 0.145. When the intermarker distance was assumed to be 15 cM, the expected maximum lod score increased from 6.12 to 6.57 and the average $\hat{z}_0$ decreased from 0.124 to 0.113. The average $\hat{d}_1$ was close to the true value of 5 cM, 4.7 and 4.9 for the 5 cM and 15 cM map, respectively. Apparently when the intermarker distance is incorrectly specified, the lod score and the estimate of the recurrence risk ratio reflect the underlying recombination fraction. For example, when the true intermarker distance is smaller than that specified in the analysis, there are fewer recombination events than expected, resulting in a higher lod score and an overestimate of the recurrence risk ratio.

DISCUSSION

The difficulties involved in the study of complex genetic disease by standard lod score methods have motivated several investigators to develop statistical techniques to map genes for such traits. Coupling ASP and affected-relative-pair methods for linkage analysis with recent advances in genotyping technology and the availability of dense genetic marker maps are currently active research areas. Several investigators have been and are pursuing methods similar to ours by using a standard linkage framework [Hyer et al., 1991] or a different parameterization of the disease model [Olson, 1995a]. In addition, similar work has been done in the context of interval mapping of quantitative trait loci using relative pairs in experimental animals [Haley and Knott, 1992] and in humans [Goldgar, 1990; Fulker and Cardon, 1994; Olson, 1995b]. A variety of genetic analysis techniques likely will be required to identify genetic components of common human diseases.

Developing a Genome Screening Strategy

Our results suggest that genotyping ASPs and no other family members with a map of genetic markers evenly spaced at 10- to 20-cM intervals provides an efficient strategy to map genes for complex genetic diseases when the limiting factor is the number of genotypes. Our simulations suggest that using a critical value of the lod score of $k = 1$ for linkage detection works well to identify intervals which may harbor a disease locus, while a critical value of $k = -2$ or even $k = -1$ for linkage exclusion works well to identify regions that may be excluded from further consid-

eration for a gene with the specified marginal effect. Exclusion mapping provides a means for setting aside intervals which may not include a disease gene, saving additional typing efforts for those intervals which seem promising, or at least, provide little evidence against linkage. Using a lod score critical value of $k = 1$ for linkage detection minimizes the risk of missing a disease locus and results in investigation of a manageable number of type I errors. In a genome scan based on typing 300 markers, we would expect about eight false-positives to be pursued. When a lod score greater than 1 is observed, the genomic region can be targeted for additional genetic studies. This investigation should include genotyping other available families or family members, genotyping additional markers in the intervals, and confirmatory analysis in independent samples. In addition, other genetic analysis methods can be employed. These might include linkage analysis with standard mode-of-inheritance based methods, combined linkage and segregation analysis, positional candidate gene studies, or disequilibrium mapping.

We emphasize that a lod score greater than 1 for linkage detection is not meant to indicate conclusive or even strong evidence for linkage [Elston, 1992]. Rather, we see this screening strategy as a means to prioritize genomic regions for further study. Intuitively, intervals with large positive lod scores should be investigated immediately, whereas areas with smaller positive lod scores may be investigated later. It also may be useful to develop a weighting scheme based on flanking marker heterozygosity when ordering intervals for further analyses. This strategy should provide an efficient strategy for rapid screening of the genome. It is important that any positive linkage results be confirmed in other samples of families when using this or any other linkage method for mapping a complex genetic trait.

Modification of this strategy may be required in studying a particular complex disease. For diseases in which finding sufficient ASPs is difficult, typing additional nuclear family members can increase information for linkage, especially when one or both parents may be genotyped. In the absence of parental genotypes, genotyping siblings in addition to the ASP provides more information for linkage than could be obtained from the ASP alone. In addition, it is useful to collect parents and additional siblings for subsequent genotyping to facilitate haplotype reconstruction and disequilibrium mapping once a disease gene is mapped. Finally, modifications in the lod score critical values may be required if families come from significantly admixed populations.

Bias of Parameter Estimates

The maximum likelihood estimates of the distance d_1 between the first flanking marker and the disease locus, and the probability that an ASP shares 0 genes identical by descent z_0, can be biased. In some cases this bias is forced by the fact that the true value of the parameter is on the boundary of the parameter space. For example, when there is no linked locus in the interval, $z_0 = 0.25$ and the estimate of z_0 must be ≤ 0.25. Similarly, when the disease locus is near a genetic marker, the estimate of d_1 is biased upward since we only considered positions within the interval in our analyses. The bias in the estimate of the disease locus position disappears when full multipoint mapping is used instead of interval mapping [Risch, 1990b; Fulker et al., 1995]. Risch [1993] has developed a full multipoint method of analysis for the case of typed parents [Risch, 1990b, 1993; Olson, 1995b]. We have extended this multipoint method to the case of untyped parents (unpublished results).

Sensitivity to Errors in the Marker Map

Our method is robust to errors in the map distance between the flanking markers, except when the errors are extreme. As with other relative pair methods when parents are not genotyped, our method can be sensitive to errors in allele frequencies. N = 200–800 ASPs provide sufficient information for reasonable estimates of allele frequencies by gene counting or maximum likelihood [Boehnke, 1991], and these estimates may then be used in the analysis. If the sample is known to be from an admixed population, the need to use a more stringent lod score criterion can be anticipated. The only difficulty in this strategy is in choosing the appropriate critical value; in fact, it may be possible to do this empirically on the basis of numbers of positive results detected as a genome scan proceeds. In any event, intervals can be ranked by lod score, and intervals with the largest lod scores can be investigated first. As mentioned previously, taking into account marker polymorphism also may be useful in ranking intervals for further analysis.

Distribution of the Maximum Lod Score

The asymptotic distribution of our maximum lod score statistic is unknown because of the nonidentifiability of θ_1 (or d_1) when $\mathbf{z} = [1/4, 1/2, 1/4]$, or for additive models when $\lambda = 1.0$. The asymptotic distribution of the usual lod score statistic multiplied by $2 \ln 10$ is a 50:50 mixture of a point mass at 0 and a χ^2 on one degree of freedom. As expected, our empirical probabilities were somewhat larger than those predicted by this mixture distribution since we are maximizing over one (λ) or two (z_0 and z_1) additional parameters. For a given data set and genotyped flanking markers, empirical power and size estimates can easily be determined by simulating several thousand replicate data sets under the null hypothesis of linkage ($\lambda > 1.0$) or no linkage ($\lambda = 1.0$).

Applying the Method to Genetic Models That Are Not Additive

Many genetic models of interest are not additive. For many complex genetic diseases, there exists little evidence of a large dominance variance or, for our purposes, a large difference in recurrence risks to sibs compared to recurrence risks to offspring, especially for small λ values [Risch, 1990c]. While we have used additivity as a simplifying assumption in most of our simulations, our recurrence risk ratio linkage method does not require this assumption. Because without the additivity assumption we use two parameters rather than one, increased critical values will be required to maintain the same rate of false-positive linkage detection. When λ is small there may be little power to detect departures from additivity. We currently are examining the relative merits of genome screening with and without the additivity assumption.

Multi-Locus Disease Models

In this work we have assumed that the marginal effect of a given disease locus can be detected, even in the presence of other genes for the disease. The ability of our method to detect marginal effects of disease genes caused by multi-locus disease model is dependent on the distortion of the ASP IBD distribution from $z = [1/4, 1/2, 1/4]$ at each locus. It may be that diseases exist for which the distortion of the marginal ASP IBD distribution is not detectable but that the distortion in the

ASP IBD distribution is detectable for combinations of loci. In that case, a genome scan may not identify single intervals as important in disease etiology, and two or more intervals will have to be examined simultaneously. We plan to examine the impact of multiple interacting loci on the genome screening strategy described above. Extension of our method to multiple disease loci is possible at the cost of substantially increased computation.

Extensions to Multiple Affected Family Members

In our method as we have described it, only the disease phenotypes of the ASP are used; disease status of other family members is ignored. When there are more than two affected siblings in a family or two or more ASPs in a pedigree, we can construct all possible ASPs and analyze the sib pairs and if desired, the complete nuclear families, as if they were independent. This violation of the assumption of the independence of the ASPs may result in a maximum lod score distribution with heavier tails than the distribution with the same number of independent ASPs, introducing an anti-conservative bias in the P values. When positive linkage results are obtained, an additional simulation step should be included to estimate the empirical P value when some of the ASPs are from the same sibships or pedigrees.

Incorporating disease status information on other relatives of the ASP is more complicated. To do so efficiently requires that we assume knowledge of a particular genetic model in order to identify the expected IBD sharing distribution among the set of affected relatives. This is a difficult problem and depends on the pedigree structure. As an intermediate step to incorporating disease phenotype on all family members we could stratify by parental disease phenotype. Risch (unpublished data) examined power for families including zero, one, or two affected parents and up to three affected siblings and found that pairs with two affected parents generally have lower power than pairs with zero or one affected parents for most genetic models examined. Sibships with zero or one affected parents exhibited similar deviations in IBD sharing and similar power to detect linkage. When ASPs are plentiful, it may be more efficient to pool sibships with zero to one affected parents, and avoid sibships with two affected parents. We plan to examine the effect of incorporating the disease phenotypes of additional nuclear family members, including stratifying families by parental disease phenotype.

CONCLUSIONS

This extension of Risch's likelihood-based mode-of-inheritance–free method of interval mapping for complex genetic diseases to use information from relatives to infer IBD status of the ASP is computationally efficient and flexible. It makes efficient use of additional siblings to increase information about marker and hence disease locus IBD status in the ASP, and it allows for disease gene localization and exclusion. The underlying genetic model based on the recurrence risk ratio is simple, and the parameter(s) are easily interpreted. New methods of automated genotyping can provide a first-pass genome scan of large numbers of families very rapidly. This statistical method allows for efficient analysis of those data. The FORTRAN computer package SIBLINK, for linkage detection, exclusion mapping, and estimation of P values, is available from the authors free of charge.

ACKNOWLEDGMENTS

Support for this work was provided by grants HG00376 (M.B.), GM52205 (S.-W.G.), and HG00348 (N.R.) from the National Institutes of Health.

REFERENCES

Bishop DT, Williamson JA (1990): The power of identity-by-state methods for linkage analysis. Am J Hum Genet 46:254–265.

Boehnke M (1991): Allele frequency estimation from data on relatives. Am J Hum Genet 48:22–25.

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SCL, Jenkins SC, Palmer SM, Balfour KM, Rowe BR, Farrall M, Barnett AH, Bain SC, Todd JA (1994): A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130–136.

Elston RC (1992): Designs for the global search of the human genome by linkage analysis. In "Proceedings of the XVIth International Biometric Conference," Hamilton, New Zealand, December 7–11, 1992, pp 39–51.

Elston RC, Stewart J (1971): A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542.

Field LL, Tobias R, Magnus T (1994): A locus on chromosome 15q26 (IDDM3) produces susceptibility to insulin-dependent diabetes mellitus. Nat Genet 8:189–194.

Freimer NB, Sandkuijl LA, Blower SM (1993): Incorrect specification of marker allele frequencies: effects on linkage analysis. Am J Hum Genet 52:1102–1110.

Fulker DW, Cardon LR (1994): A sib-pair approach to interval mapping of quantitative trait loci. Am J Hum Genet 54:1092–1103.

Fulker DW, Cherny SS, Cardon LR (1995): Multipoint interval mapping of quantitative trait loci using sib pairs. Am J Hum Genet 56:1224–1233.

Goldgar DE (1990): Multipoint analysis of human quantitative genetic variation. Am J Hum Genet 47:957–967.

Haldane JBS, Smith CAB (1947): A new estimate of the linkage between the genes for colour-blindness and haemophilia in man. Ann Eugen 14:10–31.

Haley CS, Knott SA (1992): A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324.

Hashimoto L, Habita C, Beressi JP, Delephine M, Besse C, Cambon-Thomsen A, Deschamps I, Rotter JI, Djoulah S, James MR, Froguel P, Weissenbach J, Lathrop GM, Julier C (1994): Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 371:161–164.

Hauser ER, Boehnke M, Risch N, Guo SW (1994): Affected-sib-pair interval mapping and exclusion for complex genetic traits: inferring identify by descent status from relatives. Am J Hum Genet [Suppl]55:A26.

Holmans P (1993): Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–374.

Hyer RN, Julier C, Buckley JD, Trucco M, Rotter J, Spielman R, Barnett A, Bain S, Boitard C, Deschamps I, Todd JA, Bell JI, Lathrop GM (1991): High-resolution linkage mapping for susceptibility genes in human polygenic disease: Insulin-dependent diabetes mellitus and chromosome 11q. Am J Hum Genet 48:243–257.

Lathrop GM, Lalouel JM, Julier C, Ott J (1984): Strategies for multilocus linkage analysis in humans. Proc Natl Acad Sci USA 81:3443–3446.

Morton NE (1955): Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318.

Olson JM (1995a): Multipoint linkage analysis using sibpairs: an interval mapping approach for dichotomous outcomes. Am J Hum Genet 56:788–798.

Olson JM (1995b): Robust multipoint linkage analysis: an extension of the Haseman-Elston method. Genet Epidemiol 12:177–193.

Penrose LS (1935): The detection of autosomal linkage in a data which consist of pairs of brothers and sisters of unspecified parentage. Ann Eugen 6:133–138.

Risch N (1987): Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 40:1–14.

Risch N (1990a): Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–228.

Risch N (1990b): Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–241.

Risch N (1990c): Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. Am J Hum Genet 46:242–253.

Risch N (1992): Corrections to "Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs." Am J Hum Genet 51:673–675.

Risch N (1993): Exclusion mapping for complex diseases. Am J Hum Genet [Suppl]53:#185.

Strittmatter WJ, Saunders AM Schmechel D, Pericak-Vance M, Enghild J, Salvesen GS, Roses AD (1993): Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. Proc Natl Acad Sci USA 90:1977–1981.

Suarez BK, Rice JP, Reich T (1978): The generalized sib pair IBD distribution: its use in the detection of linkage. Ann Hum Genet 42:87–94.

Weeks DE , Lange K (1988): The affected-pedigree-member method of linkage analysis. Am J Hum Genet 42:315–326.

Weeks DE, Lange K (1992): A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859–868.