Am. J. Hum. Genet. 57:439–454, 1995

Complete Multipoint Sib-Pair Analysis of Qualitative and Quantitative Traits

Leonid Kruglyak[1] and Eric S. Lander[1,2]

[1]Whitehead Institute for Biomedical Research, and [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA

Summary

Sib-pair analysis is an increasingly important tool for genetic dissection of complex traits. Current methods for sib-pair analysis are primarily based on studying individual genetic markers one at a time and thus fail to use the full inheritance information provided by multipoint linkage analysis. In this paper, we describe how to extract the complete multipoint inheritance information for each sib pair. We then describe methods that use this information to map loci affecting traits, thereby providing a unified approach to both qualitative and quantitative traits. Specifically, complete multipoint approaches are presented for (1) exclusion mapping of qualitative traits; (2) maximum-likelihood mapping of qualitative traits; (3) information-content mapping, showing the extent to which all inheritance information has been extracted at each location in the genome; and (4) quantitative-trait mapping, by two parametric methods and one nonparametric method. In addition, we explore the effects of marker density, marker polymorphism, and availability of parents on the information content of a study. We have implemented the analysis methods in a new computer package, MAPMAKER/SIBS. With this computer package, complete multipoint analysis with dozens of markers in hundreds of sib pairs can be carried out in minutes.

Introduction

Sib pairs are an increasingly important tool for genetic analysis of complex traits. Sib pairs are relatively easy to ascertain in large numbers and tend to be more closely matched for age and environment than other relative pairs. Moreover, sib-pair studies require no prior assumptions about such parameters as mode of inheritance, penetrance, phenocopy rate, and disease allele frequency.

Conceptually, sib-pair analysis is straightforward—one simply determines whether each sib pair shares 0, 1, or 2 alleles identical by descent (IBD) at a locus of interest. For a qualitative trait, affected sib pairs should share alleles IBD more often than expected under random Mendelian segregation. For a quantitative trait, sib pairs should show a correlation between the magnitude of their phenotypic difference and the number of alleles shared IBD.

In practice, the situation is more complicated because one cannot unambiguously determine the number of alleles shared IBD at every position along the genome. Most sib-pair studies have focused on studying individual markers one at a time. Such analyses are inadequate for two reasons: (i) the exact IBD status cannot always be inferred at the marker loci (the problem is most acute when parents are unavailable for study, since inferences must be drawn based only on identity-by-state (IBS) information for the offspring, but it occurs even when parents are available, since some parental mating types are not fully informative); and (ii) the IBD status at locations other than marker loci is not assessed.

Various partial solutions have been proposed over the years. For qualitative traits, a common approach has been to focus only on the subset of sib pairs for which IBD allele sharing can be determined unambiguously; this approach clearly wastes information from sib pairs that are not fully informative. Another approach is to focus solely on IBS sharing; this approach fails to take advantage of information available from multiple markers and is less powerful than analyses based on identity by descent (Bishop and Williamson 1990). Risch (1990*b*) proposed a method based on maximum likelihood (ML) that takes into account partially informative matings; a similar method was proposed by Sandkuijl (1989), and has been incorporated in the Extended Sib-Pair Analysis program. However, both approaches are based on analysis of single markers one at a time. Recently, Olson (1995) described another approach to qualitative-trait mapping that uses a pair of flanking markers; this approach requires prior knowledge of recurrence-risk information and fails to make use of additional markers. In general, interval mapping approaches that rely only on flanking markers extract all inheritance information in the case of experimental crosses between inbred strains (Lander and Botstein 1989) but fail to do

Received January 31, 1995; accepted for publication May 4, 1995.

Address for correspondence and reprints: Dr. Eric S. Lander, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142-1479.

0002-9297/95/5702-0028$02.00

so in the case of human families—owing to the fact that markers are not completely polymorphic. There is thus a clear need for a complete multipoint approach to exploit the full inheritance information in a pedigree.

For quantitative traits, the approaches have been somewhat different. Most authors have dealt with the problem of incomplete information by replacing the unknown value of the number $\check{v}_i$ of alleles shared IBD by the ith sib pair by its expected value $\check{v}_i$, given the genetic data. Originally, the expected value $\check{v}_i$ was calculated only at individual markers, using only the information for that marker (Haseman and Elston 1972). More recently, some authors have suggested calculating $\check{v}_i(s)$ at each point s in an interval based on the genotypes at the two closest flanking markers (Fulker and Cardon 1994). These ad hoc approaches, however, suffer from several limitations. First, as discussed above, examining just one or two markers does not extract the full information about IBD status, and confining analysis to two closest flanking markers is technically undesirable because it results in discontinuities in the likelihood surface. Second, the mean value $\check{v}_i(s)$ does not fully describe the IBD status at each point. Rather, one needs to know the probabilities $\pi_{i0}(s)$, $\pi_{i1}(s)$, and $\pi_{i2}(s)$ that the IBD sharing for the ith sib pair is 0, 1, or 2 at each point s. Although estimates based on the mean value $\check{v}_i$ may be asymptotically consistent in some cases, they waste information, with the result that they are statistically inefficient (that is, they have unnecessarily large variance) compared with using the full IBD distribution. Third, because the ad hoc approach is not a true ML method, it is also difficult to analyze its properties—for example, to determine appropriate significance levels. As an alternative, some authors have suggested combining single marker information so as to obtain an estimate of average IBD sharing across a chromosomal region rather than at individual points (Goldgar 1990; Guo 1994). Such approaches are an incomplete substitute for actually knowing IBD sharing at each point: they fail to exploit the full data and they cannot be used for fine-structure gene localization.

Clearly, a unified approach to sib-pair analysis of both qualitative and quantitative traits is desirable. The best solution would be a complete multipoint analysis using the information from *all* genetic markers to infer the *full* probability distribution of the IBD status at each point along the genome. Specifically, one would like to compute the probabilities $\pi_{i0}(s)$, $\pi_{i1}(s)$, $\pi_{i2}(s)$ of sharing 0, 1, and 2 alleles IBD for the ith sib pair at every point s, conditional on the data at all genetic markers along the genome. In fact, such an analysis can be easily and rapidly performed by using a linkage analysis algorithm described elsewhere by Lander and Green (1987) and recently improved by Kruglyak et al. (1995). With the IBD distribution in hand, it is possible to perform the proper ML generalization of any sib-pair method for both qualitative and quantitative traits. In the case of qualitative traits, this approach generalizes the single-point LOD score method of Risch (1990*a*, 1990*b*). For quantitative traits, methods originally considered by Haseman and Elston (1972) can be generalized and extended.

In this paper, we describe the application of this approach to sib-pair analysis and its implementation in a new computer package, MAPMAKER/SIBS. The program uses genotype information for each sib pair, together with information about parents and additional sibs where available, to infer the IBD distribution at each point along the genome. Using the IBD distribution, the program can then perform four types of sib-pair analyses: (i) *Exclusion mapping,* to test specific hypotheses about the degree of allele sharing at each location in the genome; (ii) *ML mapping,* to identify loci involved in a qualitative trait; (iii) *Information-content mapping,* to assess the extent to which the available genetic markers have extracted the full inheritance information at each location in the genome; and (iv) *Quantitative-trait-locus (QTL) mapping,* to identify loci involved in a quantitative trait, by two parametric methods and one nonparametric method. With these tools, researchers can now extract maximal information from sib-pair studies. We also explore the effects of marker spacing, marker polymorphism, and availability of parents for typing on the information content of a study. The results, presented as graphs, should assist in study design.

Results

Calculating the IBD Distribution

Consider a collection of nuclear families P_i ($i = 1, \ldots, N$), each consisting of a sib pair S_i possibly accompanied by parents and additional siblings. In the case of a qualitative trait, both sibs will be assumed to be affected. In the case of a quantitative trait, both sibs will be assumed to have been assigned a numerical phenotype. Phenotypic information about the additional relatives will be assumed to be unavailable. The pedigree members are genotyped for a collection of genetic markers with known map locations distributed throughout the genome.

The inheritance pattern of each pedigree P_i at each location s is completely specified by the inheritance vector $V_i(s)$ (Lander and Green 1987), in which each component corresponds to a particular meiosis and the coordinate is 0 or 1 according to whether the offspring inherits the allele at position s from the paternally or maternally derived chromosome. For a sib pair, the vector has four components: two for each sibling, with one specifying whether the mother's maternally or paternally derived allele has been transmitted to the sib and the

other specifying the corresponding information for the father. Typically, the actual inheritance vector $V_i(s)$ cannot be uniquely determined from the genetic marker data. Instead, one can calculate the probability distribution for $V_i(s)$ conditional on the genetic marker data. Lander and Green (1987) described a hidden-Markov-model algorithm for computing the probability distribution over $V_i(s)$, which has recently been improved by Kruglyak et al. (1995). The improved algorithm has recently been implemented for nuclear pedigrees and makes feasible complete multipoint homozygosity mapping in minutes (Kruglyak et al. 1995). Applying the same approach to the present situation, the algorithm allows rapid calculation of the probability distribution over the inheritance vectors $V_i(s)$ at each location in the genome. Given this information, one need only add the probabilities of the appropriate inheritance vectors to calculate the probabilities $\pi_{i0}(s)$, $\pi_{i1}(s)$, and $\pi_{i2}(s)$ that the ith sib pair shares 0, 1, or 2 alleles IBD at position s.

As with all current multipoint approaches, the algorithm neglects effects of crossover interference. In fact, the effect of interference on detection of linkage is quite minor (Terwilliger and Ott 1994). If desired, interference effects can be partially included through a choice of map function.

MAPMAKER/SIBS Computer Package

We wrote a computer package, MAPMAKER/SIBS, that implements this approach to calculate the IBD distribution $\pi_{i0}(s)$, $\pi_{i1}(s)$, and $\pi_{i2}(s)$ and then uses it to calculate various sib-pair statistics discussed below. Figure 1 illustrates an example of the IBD distribution for a single sib pair without parents, genotyped across an entire chromosome. In the example shown, there are apparent crossovers between 30 and 40 cM (from 2-sharing to 1-sharing) and between 70 and 80 cM (from 1-sharing to 0-sharing).

The program is extremely rapid. Analysis of 200 sib pairs with chromosomal map containing 50 genetic markers required <4 min on a DEC 3000 Alpha work station. The speed is essentially independent of the presence or absence of parental information. The program is written in C and is freely available from the authors by anonymous ftp (at ftp-genome.wi.mit.edu, in the directory distribution/software/sibs) or from our World Wide Web site (http://www-genome.wi.mit.edu/ftp/distribution/software/sibs). It takes two files as input: (i) a pedigree file, specifying individuals in each pedigree P_i and their genotypes; and (ii) a map file, specifying the locations and allele frequencies of the genetic markers. The appropriate file formats are described in an accompanying user's manual; files in LINKAGE format (Terwilliger and Ott 1994) may be used directly. The program handles autosomal and sex-linked data.

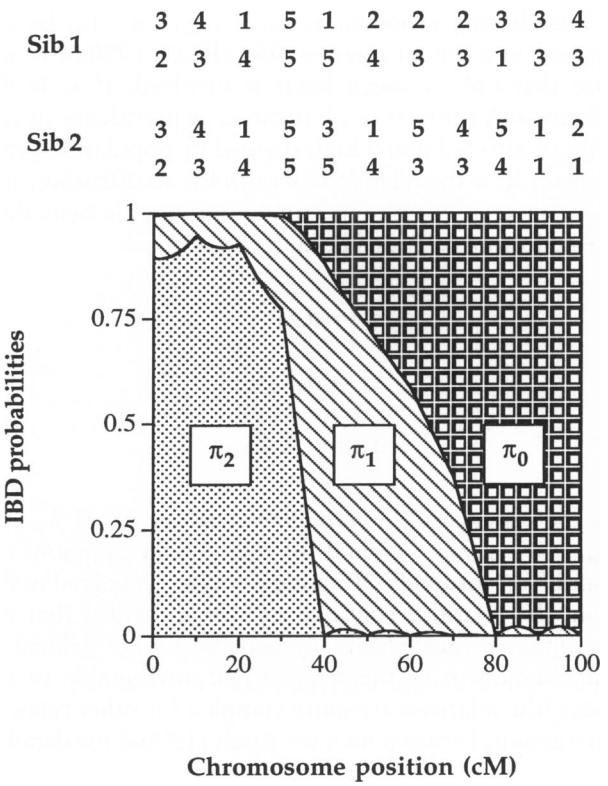

Figure 1 The IBD distribution for a single sib pair without available parents. The sibs have been genotyped at markers spaced every 10 cM along a chromosome of total length 100 cM. The markers are assumed to have five equally frequent alleles, corresponding to a heterozygosity of .8. Shaded areas represent the probabilities π_0, π_1, and π_2 of sharing 0, 1, and 2 alleles IBD, respectively. Marker genotypes for the two sibs are shown above the graph. Crossovers appear to have occurred between 30 and 40 cM (from 2-sharing to 1-sharing) and between 70 and 80 cM (from 1-sharing to 0-sharing).

Qualitative Traits: Characterizing IBD Sharing

For a qualitative trait, each locus in the genome can be characterized by the expected proportions z_0, z_1, and z_2 of affected sib pairs sharing 0, 1, and 2 alleles IBD. For a locus involved in causing the trait, (z_0, z_1, z_2) will differ from the expected Mendelian proportions (α_0, α_1, α_2) = ($\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$). Biological consistency constrains (z_0, z_1, z_2) to lie within the "possible triangle" (Holmans 1993) defined by

$$z_0 + z_1 + z_2 = 1; \quad z_1 \leq \tfrac{1}{2}; \quad z_1 \geq 2z_0 . \qquad (1)$$

The further assumption of no dominance variance is equivalent to the constraint:

$$z_1 = \tfrac{1}{2}, \qquad (2)$$

in which case, the sharing proportions are described by the single parameter z_2.

The sharing proportions (z_0, z_1, z_2) can also be expressed in terms of relative risks (Risch 1990*a*), in the case that only a single locus is involved. If λ_S is the relative-risk ratio for a sib (defined as prevalence in siblings of affected individuals divided by population prevalence), λ_O is the relative-risk ratio for an offspring, and λ_M is the relative-risk ratio for a monozygotic twin, then the following relations hold:

$$z_0 = \alpha_0/\lambda_S ;$$
$$z_1 = \alpha_1\lambda_O/\lambda_S ; \quad (3)$$
$$z_2 = \alpha_2\lambda_M/\lambda_S .$$

In the absence of dominance variance, $\lambda_O = \lambda_S$ and $\lambda_M - 1 = 2(\lambda_S - 1)$, and the relations in equations (3) simplify accordingly. If multiple loci are involved in the trait, the relations continue to hold, provided that the loci interact multiplicatively and the λ's are defined as the component of the relative risk attributable to the locus (the relations are more complex for other types of interactions between loci; see Risch [1990*a*] for details).

Qualitative Traits: Exclusion Mapping

In scanning the genome, one may wish to identify and exclude those regions unlikely to have a major effect on the trait—for example, to contain a locus with a relative risk exceeding a given threshold. One may then focus attention on the remainder of the genome.

Such exclusion mapping is easily performed by calculating a LOD score comparing the likelihood of the data arising under any specific hypothesis (z_0, z_1, z_2) to the likelihood under the Mendelian alternative (α_0, α_1, α_2). By Bayes's theorem, the probability $\rho_{ij}(s)$ of observing the marker genotype data for the ith sib pair, given that the pair shares j alleles IBD at position s, is proportional to $\pi_{ij}(s)/\alpha_j$. The likelihood ratio for the ith sib pair at position s is thus

$$L_i(s) = \frac{z_0\rho_{i0} + z_1\rho_{i1} + z_2\rho_{i2}}{\alpha_0\rho_{i0} + \alpha_1\rho_{i1} + \alpha_2\rho_{i2}} ,$$

and the LOD score for the collection of pedigrees is

$$Z(s) = \sum_i \log_{10}L_i(s) .$$

Using MAPMAKER/SIBS, one can test any specific collection of hypotheses—specified either in terms of the sharing proportions z_0, z_1, z_2 (for any model) or in terms of the relative risks λ_S, λ_O, and λ_M (for single locus or multiplicative models). An exclusion map is given in figure 2. The example shows the LOD score for 100 sib pairs without parents along two representative chromosomes—the first containing no loci affecting the trait and the second containing a locus at 25 cM with no dominance variance and a relative risk $\lambda_S = 5$. The LOD score is shown for the hypotheses $\lambda_S = 1.2$, 1.5, 2, 3, 5, and 10 for the locus, assuming a multiplicative model and no dominance. Adopting the traditional exclusion criterion of $Z < -2$, it is possible to exclude a locus with $\lambda_S \geq 3$ from the entire first chromosome.

Qualitative Traits: ML Mapping

In searching for loci involved in a complex trait, one needs to scan the genome to identify regions of significant excess allele sharing. A simple and effective approach is to estimate the ML values of the allele-sharing proportions ($\hat{z}_0$, $\hat{z}_1$, $\hat{z}_2$) at each location along the genome and then compute a maximum LOD score $\hat{Z}(s)$ at each location, comparing the likelihood of the observed data arising under these ML values to the likelihood under random Mendelian segregation.

MAPMAKER/SIBS allows one to calculate the ML proportions ($\hat{z}_0$, $\hat{z}_1$, $\hat{z}_2$), either subject only to the "possible triangle" constraint (eq. [1]) or to the additional constraint of no dominance variance (eq. [2]). The constrained maximizations are performed by a simple expectation-maximization (EM) algorithm, described in appendix A.

Figure 3C shows a LOD score plot for the same data set used for the exclusion map in figure 2. Note that the LOD score is never negative, because the ML solution ($\hat{z}_0$, $\hat{z}_1$, $\hat{z}_2$) at each location can never be worse than the random Mendelian segregation. The LOD score near the true locus on the second chromosome reaches a peak of 6.8 within 1 cM of the true locus.

What LOD score threshold should be used for a genomewide false-positive rate of 5% in a genomewide search? The correct dense-map threshold for the possible triangle method is 4.0, which corresponds to a nominal (single-test) significance level of 2×10^{-5} (Lander and Schork 1994). Holmans (1993) correctly observed that a LOD score of 2.3 corresponds to a nominal significance level of 10^{-3} but did not discuss whether this threshold is adequate for a genomewide false-positive rate of 5%. In fact, a LOD score of $\geq$2.3 will occur by chance (i.e., when there are no susceptibility loci in the genome) *somewhere* in the genome, with ~80% probability. Misinterpreting statements about single-test p values as statements about genomewide significance can thus lead to unacceptably high false-positive rates.

Although the threshold of 4.0 is based on the assumption of a dense map, we strongly recommend it even for studies using a map of moderately spaced markers. The

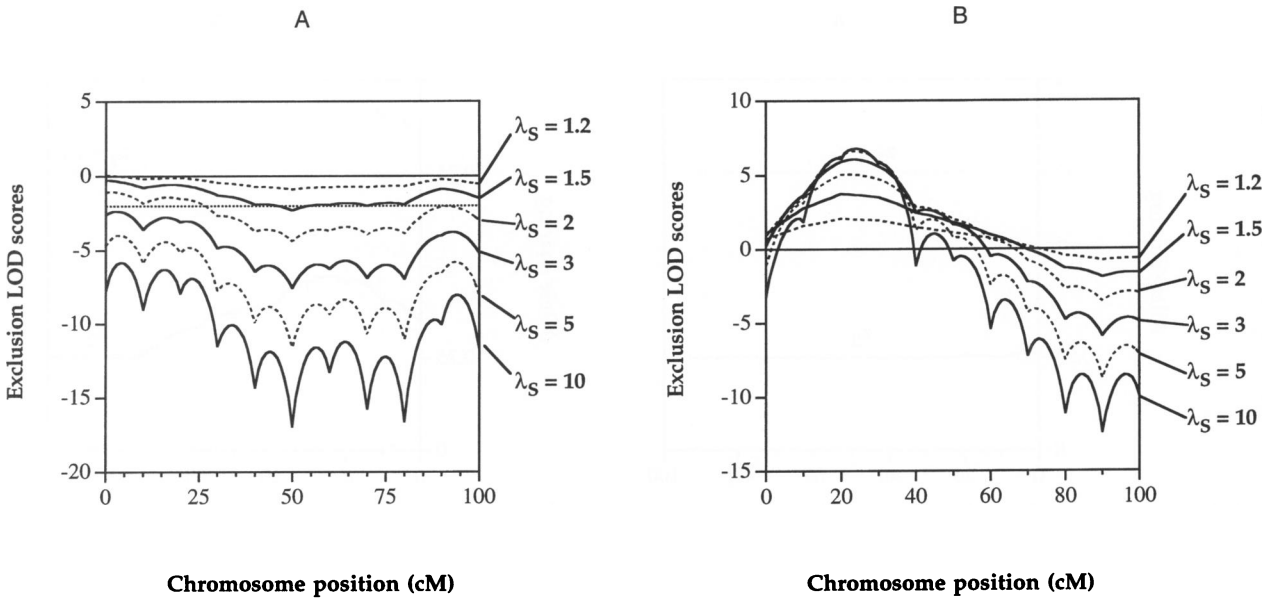

Figure 2 Exclusion mapping. Results of a simulation in which 100 sib pairs without available parents are genotyped for markers spaced every 10 cM on a 100 cM chromosome, each with five equally frequent alleles, corresponding to a heterozygosity of .8. LOD scores are computed under the assumptions of no dominance variance and a single locus (or several loci interacting multiplicatively) for the hypotheses λ_S = 1.2, 1.5, 2, 3, 5, and 10 (in the case of several loci, λ_S is the risk ratio associated with one locus). *A,* The chromosome contains no loci affecting the trait. It is possible to exclude the hypothesis of a locus with $\lambda_S \geq 3$ from the entire chromosome at the exclusion threshold of LOD < -2 (*dotted horizontal line*). *B,* A locus with $\lambda_S = 5$ and no dominance variance is located at 25 cM. High values of λ_S can be excluded from a large region at the other end of the chromosome, while all λ_S values give positive LOD scores at 25 cM.

reasons are twofold: (1) Although the threshold will be somewhat more stringent than the 5% significance level, the difference is relatively small. (2) It can be anticipated that investigators will increase the marker density in regions showing suggestive evidence of linkage, and that other investigations are likely to carry out similar studies of the same sample or, at least, of the same trait. The dense-map assumption is thus a fair description of the situation. Because of the selection bias that only positive results tend to be reported, the scientific literature must regard all proposed linkages as if they were obtained by scanning the entire genome with a dense map. Otherwise, published papers will contain an excess of false-positive linkages. This recommendation is consistent with the traditional rule that a LOD threshold of 3.0 be required for linkage analysis of human monogenic traits, regardless of the number of chromosomes or marker loci examined. The only difference is the threshold of 4.0, which is based on the appropriate calculation of the genomewide false-positive rate for sib-pair studies. It may be worth reporting studies in which the observed p value falls short of the nominal 2×10^{-5} significance level, but such results should be considered "suggestive" since they would be expect to occur simply by chance in >5% of studies. Such suggestive results may, of course, become statistically significant when one pools the data from the initial study with the data from subsequent studies.

Information-Content Mapping

How dense a genetic map should be used in sib-pair analysis? Ideally, one would like to know the IBD status with certainty at every location in the genome. Unfortunately, this would require an infinitely dense map with infinitely polymorphic markers. Instead, it is sensible to ask how close one has come to extracting the full information by using a given collection of genetic markers.

Using the methods above, one can obtain a good measure of the amount of IBD information extracted. The measure is based on the variance of the IBD distribution. Before genotyping is performed, the a priori IBD distribution for the ith sib pair is ($^1/_4$, $^1/_2$, $^1/_4$) for sharing 0, 1, or 2 alleles IBD. This distribution has mean 1 and variance $\sigma^2_{i,\,\text{initial}} = {}^1/_2$. After genotyping has been performed, the a posteriori IBD distribution at a point s is calculated to be $(\pi_{i0}(s), \pi_{i1}(s), \pi_{i2}(s))$. This new distribution has a new variance $\sigma^2_{i,\,\text{residual}}(s)$ at position s. If the IBD sharing at s is known with certainty, $\sigma^2_{i,\,\text{residual}}(s) = 0$. The quantity

$$r^2(s) = 1 - \frac{\Sigma_i \sigma^2_{i,\,\text{residual}}(s)}{\Sigma_i \sigma^2_{i,\,\text{initial}}} = 1 - 2\langle \sigma^2_{\text{residual}}(s) \rangle ,$$

where $\langle \ldots \rangle$ denotes the average over all the sib pairs, measures the extent to which the uncertainty in the

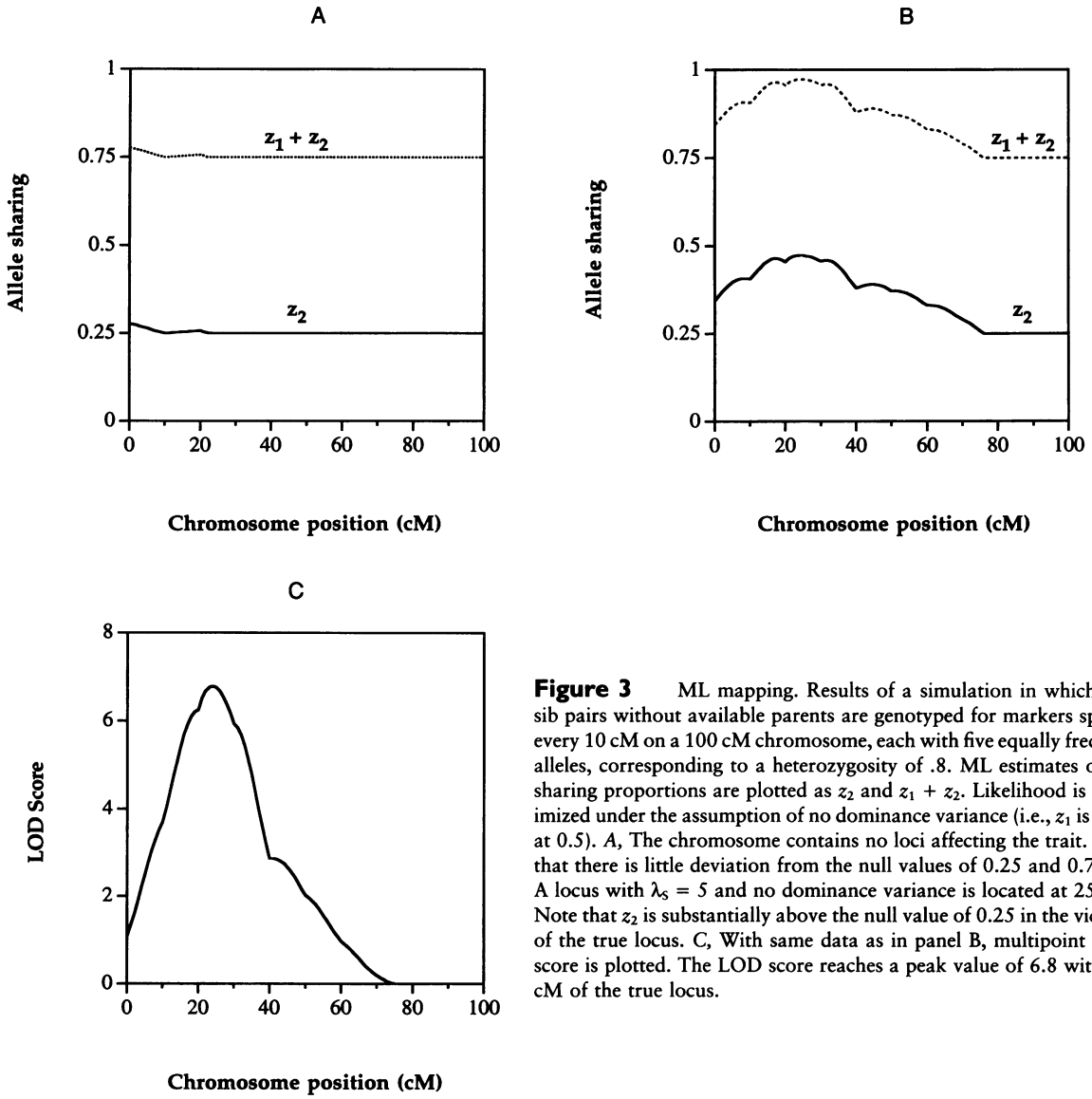

Figure 3 ML mapping. Results of a simulation in which 100 sib pairs without available parents are genotyped for markers spaced every 10 cM on a 100 cM chromosome, each with five equally frequent alleles, corresponding to a heterozygosity of .8. ML estimates of the sharing proportions are plotted as z_2 and $z_1 + z_2$. Likelihood is maximized under the assumption of no dominance variance (i.e., z_1 is fixed at 0.5). *A*, The chromosome contains no loci affecting the trait. Note that there is little deviation from the null values of 0.25 and 0.75. *B*, A locus with $\lambda_S = 5$ and no dominance variance is located at 25 cM. Note that z_2 is substantially above the null value of 0.25 in the vicinity of the true locus. *C*, With same data as in panel B, multipoint LOD score is plotted. The LOD score reaches a peak value of 6.8 within 1 cM of the true locus.

IBD distribution for the collection of sib pairs is resolved by the genotype information. Values of $r^2(s)$ near 1 indicate that most of the inheritance information has been extracted, while values substantially below 1 indicate that little information has been extracted and that additional markers would be useful. (We note that it is formally possible that the information content $r^2(s)$ can be negative. As a simple example, consider the case of a single family studied with a single marker at which both parents and both sibs have heterozygous genotypes a/b. The IBD distribution at the locus is then $(\pi_0, \pi_1, \pi_2) = (1/2, 0, 1/2)$, which has greater variance than the a priori distribution $(1/4, 1/2, 1/4)$. The negative information content correctly reflects our increased uncertainty about the extent of IBD sharing. In situations with multiple sib pairs and multiple markers, negative information contents are unlikely to be encountered in practice.)

Information-content mapping is illustrated in figure 4. The examples show results for simulations of 100 sib pairs analyzed with genetic maps of various densities and polymorphism rates. Observe that a genetic map with an average spacing of 10 cM and marker heterozygosity of .8 (a polymorphism rate typical of the current generation of human microsatellite polymorphisms) extracts ~70% of the total information about IBD status at marker loci and ~60% of the total information at the middle of intervals.

Information-content mapping makes clear where the vast majority of the IBD information has been extracted and focuses attention on the regions in which further genotyping would provide substantial additional infor-

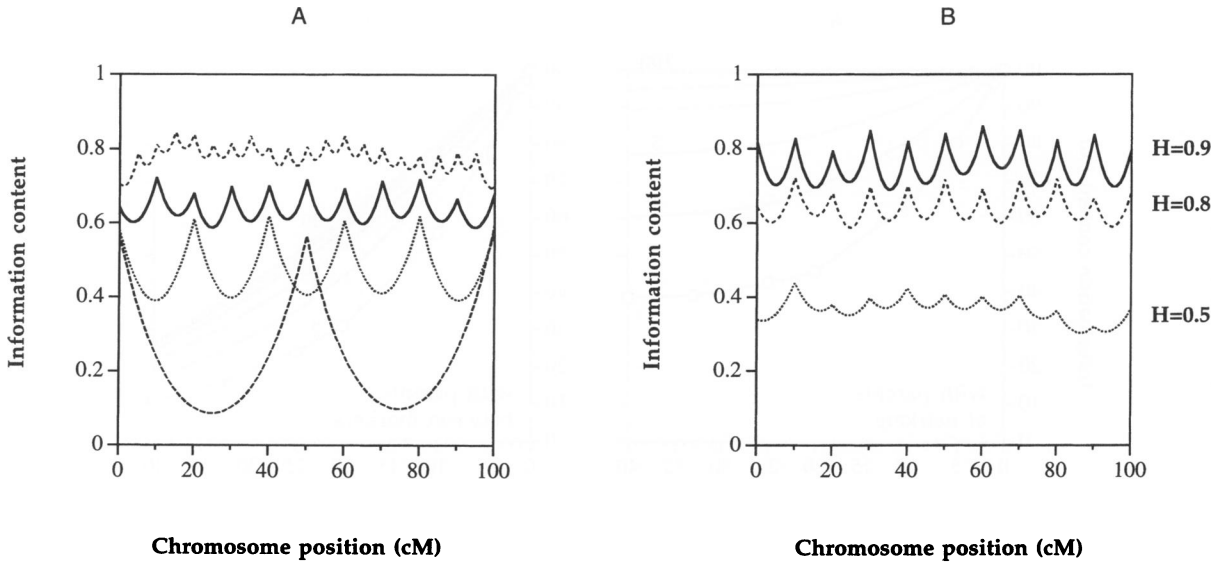

Figure 4 Information-content mapping. The amount of IBD information extracted by the genotype data ($r^2(s)$; see text for details) is plotted along a chromosome for different marker densities and heterozygosities, for a collection of 100 sib pairs without available parents. *A*, Effect of marker density. Markers with five equally frequent alleles, corresponding to a heterozygosity of .8, are spaced every 5 cM, 10 cM, 20 cM, and 50 cM. Note that the use of a denser map increases the information extracted, not only between markers but also at the marker locations; this illustrates the advantage of a multipoint approach over studying single markers individually. *B*, Effect of marker heterozygosity. Markers with 10, 5, and 2 equally frequent alleles (heterozygosities H = .9, .8 , and .5, respectively) are spaced every 10 cM.

mation—for example, to increase or decrease the LOD score in a region with a suggestive result.

Information Content and Study Design Considerations: The Quality of a Map

In order to plan a sib-pair study, it is useful to know the expected information content that can be extracted as a function of the study design. We performed simulations to explore the effects of map density, marker polymorphism rate, and availability of parents for genotyping. We generated 100 replicate data sets containing 100 sib pairs each, both with and without parental genotype information. The average information content for the replicates was computed as a function of marker density (assuming evenly spaced markers) and marker polymorphism rate (assuming a given number of equally frequent alleles; see fig. 5 legend for details). Information content was measured in two places: at marker loci, where it is greatest; and midway between markers, where it is lowest.

We first consider the effect of marker density, assuming a polymorphism rate of 75% representative of the current generation of microsatellite markers (fig. 5). A 10-cM map extracts ~85% of IBD information at markers and 70% midway between markers when parents are available, and 65% and 55% when parents are unavailable. With a 20-cM map, these numbers fall somewhat, to 75% and 50% when parents are available and 55% and 35% when parents are unavailable. With marker spacing >40 cM, there is essentially no interaction between markers, and, thus, no additional benefit is derived from using a map. In practice, it would be best to use a nested screening approach. For example, one might initially screen a sib-pair collection with a sufficient marker density to extract 50%–75% of the information and then employ a denser set of markers in any region showing a LOD score >1.

We next consider the tradeoff among marker polymorphism rates (fig. 6). For example, one can extract ~75% of the information between markers either by using a 9-cM map of current microsatellites (75% heterozygosity) or a 5-cM map of perfect biallelic markers (50% heterozygosity), when parents are available. More generally, the sib-pair information extracted by a microsatellite map of a given density can be equivalently obtained by using biallelic markers at a density that is higher by a factor of about two. This tradeoff is relevant to consider, inasmuch as it may be possible to achieve considerably higher degrees of automation for biallelic markers (requiring only plus-minus detection systems) than for microsatellites (requiring length measurement).

Application to Type I Diabetes (IDDM)

To test the methods described above on actual data, we applied them to a sib-pair study of IDDM recently reported by Davies et al. (1994). IDDM is known to have a substantial but complex genetic component, with the strongest genetic contribution arising from HLA on

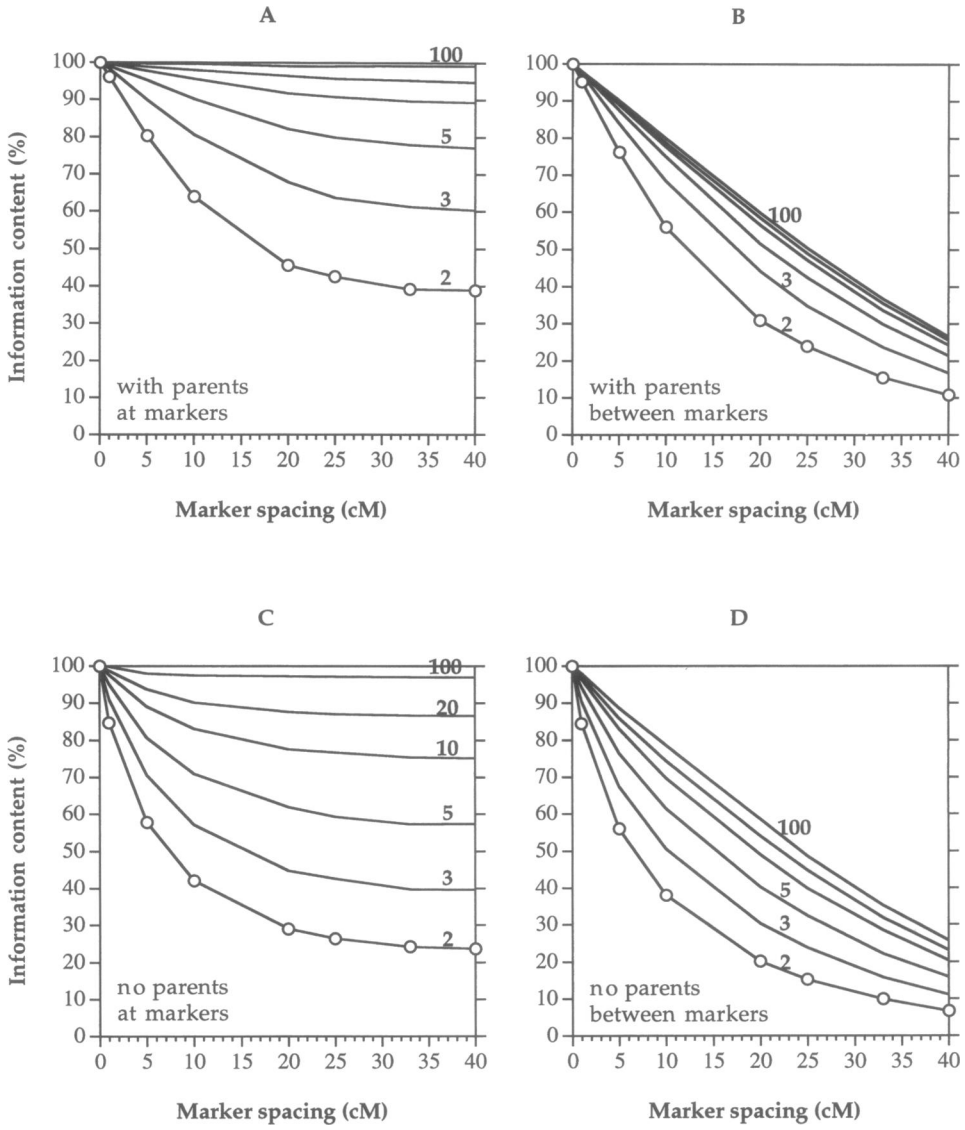

Figure 5 Expected IBD information content as a function of marker density and marker heterozygosity. A total of 100 simulations consisting of 100 sib pairs each, with and without available parents, were generated for even marker spacings of 1, 5, 10, 20, 25, 33, and 40 cM, and for markers with m = 2, 3, 5, 10, 20, and 100 equally frequent alleles. Information content was computed both at marker loci (highest information) and midway between markers (lowest information). Curves show average IBD information content for, going from top to bottom, m = 100, 20, 10, 5, 3, and 2 (the values of m are shown where space allows). For clarity, the actual data points (*circles*) are shown only for m = 2; standard deviation of the mean is not shown, as it is smaller than the size of the circles. *A,* Expected information content at markers; parents available for typing. *B,* Expected information content midway between markers; parents available for typing. *C,* Expected information content at markers; parents unavailable for typing. *D,* Expected information content midway between markers; parents unavailable for typing.

chromosome 6. Using data from chromosome 6 shared by J. L. Davies and colleagues, we carried out exclusion mapping, ML mapping, and information-content mapping. Using single-point analysis, Davies et al. (1994) previously found LOD scores of 8, 1.8, and 1.2 near HLA on 6p, near ESR on 6q (designated IDDM5), and at D6S264 on distal 6q. With multipoint ML analysis, HLA is easily detected with a peak LOD score of 11 (see fig. 7*A:* HLA marker TNFa is at 29 cM; ESR is at 128 cM; and D6S264 is at 155 cM). The region near ESR shows a peak LOD of 2.25, with a smaller peak of 1.1 near D6S264. These results illustrate the greater power of the multipoint approach—at HLA, the multipoint LOD score is 3 log units higher (corresponding to 1,000-fold greater odds in favor of linkage) than the single-point LOD score, with a smaller increase at ESR. When likelihood maximization is carried out under the assumption of no dominance variance, the peak LOD score near HLA is only 8.6, significantly lower than the unconstrained peak (fig. 7*A*). This supports

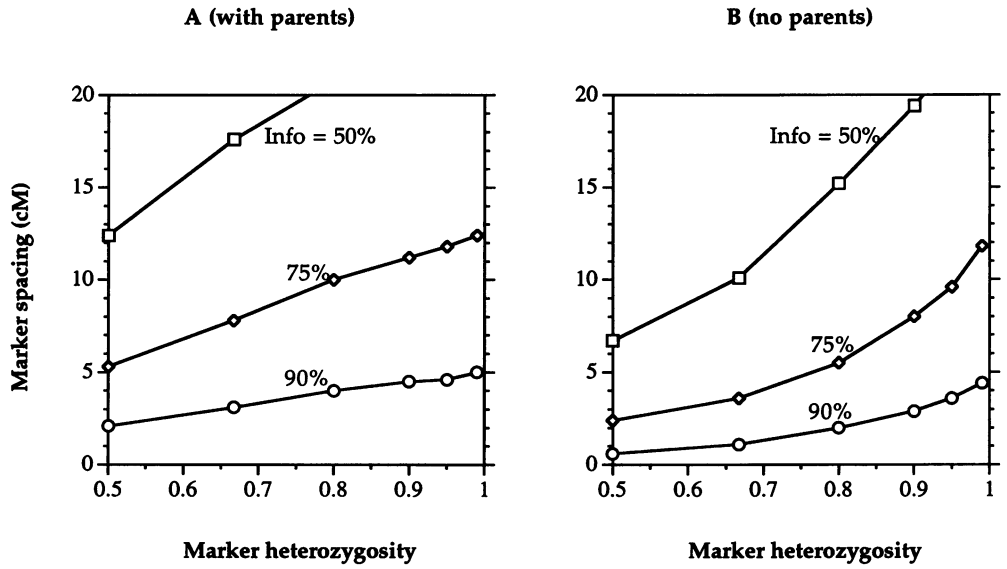

Figure 6 Map density required to achieve a given expected IBD information content. Simulations were carried out as described in fig. 5. Marker spacing in centimorgans required to extract 50% (*squares*), 75% (*diamonds*) and 90% (*circles*) of the IBD information midway between markers is plotted against marker heterozygosity. Data points are for markers with 2, 3, 5, 10, 20, and 100 equally frequent alleles (heterozygosities of .5, .67, .8, .9, .95, and .99, respectively). *A*, Parents available for typing. *B*, Parents unavailable.

the well-known partial dominance effect at HLA. No dominance effect is observed at ESR.

Exclusion mapping allows loci with substantial effects to be excluded from the middle section of chromosome 6, but only $\lambda_S = 10$ can be excluded from the entirety of the chromosome apart from a 45-cM region around HLA (fig. 7*B*). Observe that allele sharing in the HLA and ESR regions is substantially above the Mendelian expectation (fig. 7*C*). Allele sharing can be used to estimate risk ratios according to the formulas above (under the single locus or multiplicative model). At HLA, we find $\lambda_S = 2.55$ and $\lambda_O = 1.63$, once again showing evidence of dominance variance (fig. 7*D*). Note that while the best estimates of the risk ratios are higher ~10–20 cM away from HLA, the likelihood of a locus at that location is significantly lower than at HLA (see fig. 7*A*). At ESR, we obtain $\lambda_S \approx \lambda_O \approx 1.8$, consistent with no dominance. Finally, we looked at information content (fig. 7*E*), which is high around HLA (97%) and ESR (96%) and drops to ~50% in the middle of the chromosome, where the markers are less dense and no substantial excess sharing is observed.

This example illustrates the utility of the methods described in this paper. Higher LOD scores are obtained with the multipoint analysis, indicating greater power. In this study, the parents of all sib pairs were available for typing. In studies without parents, the increase in power of multipoint analysis over single-point analysis will be even greater. Continuous plots of LOD scores, allele-sharing proportions, and risk-ratio estimates provide additional insight into the data. Information-content mapping allows one to choose additional markers where they are needed most.

Quantitative Traits: Methods, Given Complete Data

We next consider sib-pair studies aimed at detecting linkage to a quantitative trait, a problem first considered by Penrose (1938). The underlying idea is simple: siblings that share more alleles at a locus affecting the trait should be more similar in phenotype than siblings that share fewer alleles. More precisely, let ϕ_{1i}, ϕ_{2i} denote the phenotypes of the two siblings; let $D_i = \phi_{1i} - \phi_{2i}$ denote the phenotypic difference; and let ν_i denote the number of alleles shared IBD at a locus of interest. At all loci in the genome, the expected value of D is 0 independent of ν (because the order of the sibs is arbitrary). At loci affecting a quantitative trait, however, the *variance* of D depends on ν—specifically, it is smaller when more alleles are shared. QTL mapping in humans, thus, corresponds to testing whether $\sigma_0^2 > \sigma_1^2 > \sigma_2^2$, where σ_j^2 denotes the variance of the difference D when j alleles are shared. (By contrast, QTL mapping in experimental crosses involves comparing *means* of progeny inheriting specific parental alleles—which is simpler and more powerful than comparing variances of phenotypic differences.)

Three basic approaches can be used to test whether $\sigma_0^2 > \sigma_1^2 > \sigma_2^2$. Below, we focus on the situation in which allele sharing is unambiguously known. In the next section, we discuss the generalization to the situation of incomplete data.

1. *Traditional Haseman-Elston QTL regression analysis.*—

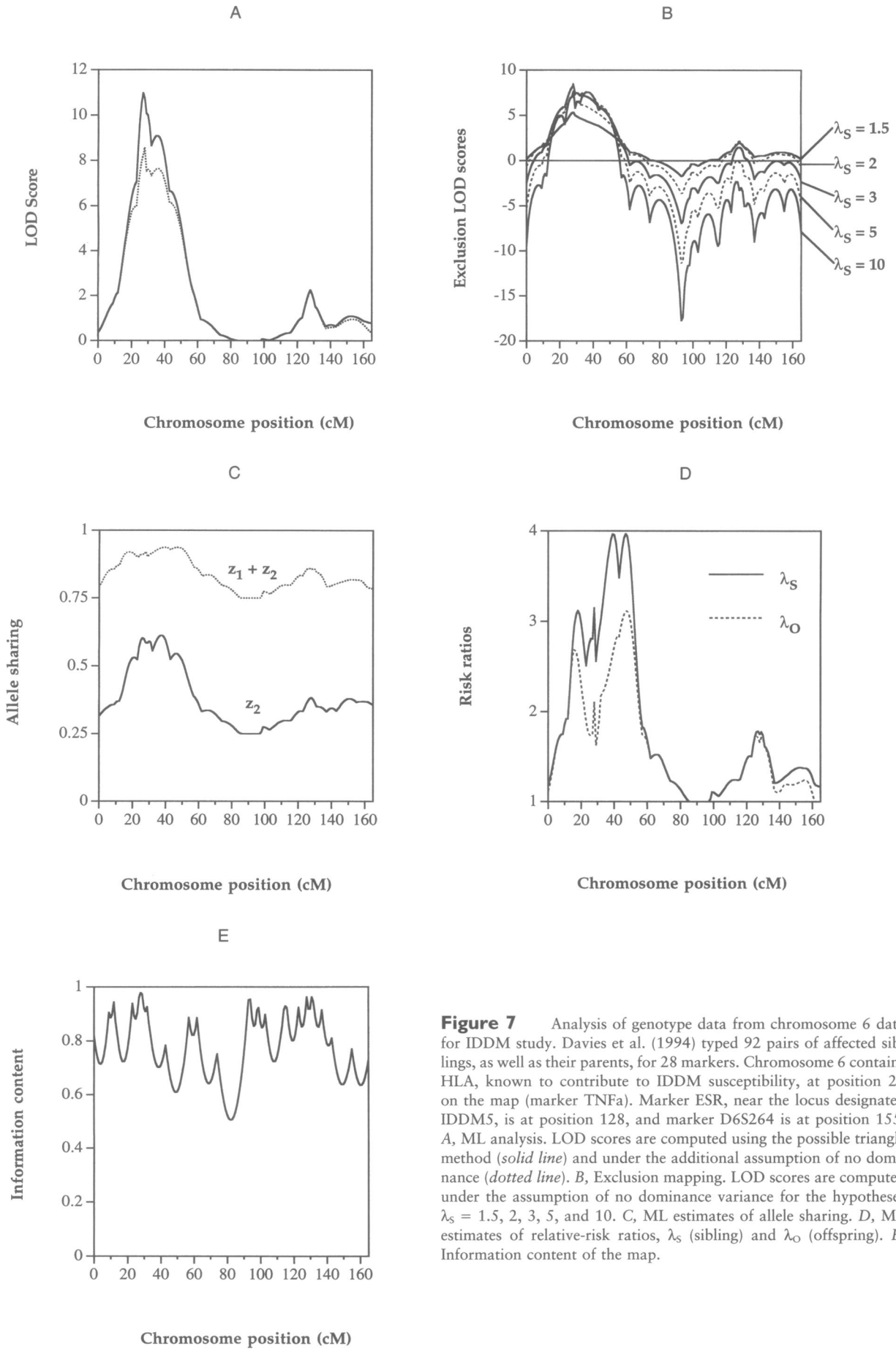

Figure 7 Analysis of genotype data from chromosome 6 data for IDDM study. Davies et al. (1994) typed 92 pairs of affected siblings, as well as their parents, for 28 markers. Chromosome 6 contains HLA, known to contribute to IDDM susceptibility, at position 29 on the map (marker TNFa). Marker ESR, near the locus designated IDDM5, is at position 128, and marker D6S264 is at position 155. *A,* ML analysis. LOD scores are computed using the possible triangle method (*solid line*) and under the additional assumption of no dominance (*dotted line*). *B,* Exclusion mapping. LOD scores are computed under the assumption of no dominance variance for the hypotheses λ_S = 1.5, 2, 3, 5, and 10. *C,* ML estimates of allele sharing. *D,* ML estimates of relative-risk ratios, λ_S (sibling) and λ_O (offspring). *E,* Information content of the map.

Haseman and Elston (1972) approached the problem of estimating variances by noting that

$$E(D_i^2 | \nu_i) = \alpha - \beta \nu_i ,$$

where $\beta = \sigma_g^2$ is the additive genetic variance attributable to the locus; a somewhat more complex relation holds when dominance effects are present. They suggested estimating β by linearly regressing D_i^2 on ν_i, which is easy to do, provided that the ν_i's are unambiguously known.

Although often used, this approach has certain statistical drawbacks. Specifically, linear regression guarantees an ML estimate only if the noise process is normally distributed and uncorrelated with the dependent variable—assumptions that are both likely to be false in the case at hand. Even if the underlying phenotype ϕ has normally distributed and uncorrelated error, the square difference D^2 does not have these properties. As a consequence, significance levels—either single-point or genomewide—cannot be evaluated by standard tests. In particular, the usual t-test formed by dividing β by its standard error is not correct, because both the estimate of the standard error and the distribution of the test statistic are based on the assumption of normal, uncorrelated error.

2. *ML QTL variance estimation.*—Alternatively, given the assumption of normally distributed phenotypic noise, one can simply derive direct ML estimates of the σ_j^2 based on the phenotypic differences observed for each value of ν (See appendix C). Estimates can be made subject to a simple constraint of biological plausibility,

$$\sigma_0^2 \geq \sigma_1^2 \geq \sigma_2^2 ,$$

or subject to the more restrictive assumption of no dominance variance:

$$\sigma_1^2 = \frac{\sigma_0^2 + \sigma_2^2}{2} .$$

Although this approach offers conceptual and technical advantages over the regression approach, it has not been adopted, probably owing to difficulties in coping with incomplete data. We show below how to overcome this problem.

3. *Nonparametric QTL analysis.*—Both approaches above require restrictive assumptions about the distribution of phenotypic effects, which may be violated in real applications. An alternative is to use a nonparametric approach, in which no assumptions are made about the distribution of phenotypic differences (Haseman and Elston 1972). Kruglyak and Lander (1995) have recently developed such an approach for genomewide QTL mapping in experimental crosses based on the Wilcoxon rank-sum test, which carries over directly to the case of human sib pairs.

In brief, the sib pairs are first ranked according to the absolute value of the phenotypic difference, with rank(i) denoting the rank of the ith sib pair. For each location s in the genome, one defines the statistic:

$$X_W(s) = \sum_{i=1}^{n} \text{rank}(i) \times f(\nu_i) ,$$

where $f(\nu)$ is a simple function of the number of alleles shared IBD, chosen to have expected value 0. In the absence of dominance effects, for example, an appropriate choice is $f(2) = -1$, $f(1) = 0$, $f(0) = 1$. (See Kruglyak and Lander [1995] concerning other choices.) In the absence of linkage, the statistic $X_W(s)$ has expectation 0 and variance $V = \frac{n\,(n+1)(2n+1)}{12}$, and the ratio $Z(s) = \frac{X_W(s)}{\sqrt{V}}$ is asymptotically distributed as a standard normal and follows an Ornstein-Uhlenbeck diffusion process (Kruglyak and Lander 1995). This property allows one to compute the genomewide significance levels for observed deviations (Lander and Botstein 1989; Kruglyak and Lander 1995). In the case of QTL mapping with sib pairs, a one-sided test ($X_W(s) > 0$) is appropriate, and the threshold for a genomewide significance level of $p = .05$ is $Z = 4.1$; this threshold corresponds to a nominal single test p value of 2×10^{-5}. Nonparametric analysis has the virtue of being robust to violations of the assumptions of normality but may be somewhat less powerful when the assumptions are satisfied.

QTL Mapping: Incomplete Information

The three approaches above are straightforwardly applied when the IBD sharing is unambiguously known, but they have not been successfully generalized to the (usual) situation of incomplete information. We show below how to generalize them by using the complete IBD distribution $\pi_0(s)$, $\pi_1(s)$, $\pi_2(s)$.

1. *Traditional Haseman-Elston QTL regression analysis.*—When the ν_i are not known with certainty, standard regression cannot be performed. Several authors have suggested performing regression with the expected values $\hat{\nu}_i$'s used in place of the unknown ν_i. Haseman and Elston (1972) describe using the expectation based only on the data for a single marker, while Fulker and Cardon (1994) generalize this by using the expectation based on the data for two flanking markers. These approaches have two drawbacks: they do not make full use of available genotype information, and they fail to carry out missing-value regression correctly.

A better approach is to perform regression using the full information inherent in the complete IBD distribu-

tion. In fact, this can be done by employing well-established ML techniques for missing-value problems (Little and Rubin 1987). In particular, one can use the EM algorithm to perform regression not on the expected value $\hat{v}_i$ but on the actual *distribution* of v_i (i.e., π_{i0}, π_{i1}, π_{i2}). Indeed, this approach has already been implemented for QTL mapping in experimental crosses in the MAPMAKER/QTL computer package. The details are given in appendix B.

2. *ML QTL variance estimation.*—In a similar fashion, the EM algorithm can be used to calculate ML estimates of the σ_i^2 even when the v_i's are not known with certainty. The details are given in appendix C.

3. *Nonparametric QTL analysis.*—In the case of nonparametric analysis, the statistic $X_W(s)$ can be replaced by its expectation over the IBD distribution: $Y_W(s) = E[X_W(s) \mid \text{data}]$. In the case of no dominance variance, this amounts to simply replacing $f(v_i)$ by its expected value at position s, which is $\pi_{i0}(s) - \pi_{i2}(s)$. $Y_W(s)$ still has expected value 0, but its variance $V(s)$ is now a function of s (here, the expectation and the variance are computed over possible realizations of the data). $Z(s) = Y_W(s)/\sqrt{V(s)}$ is asymptotically distributed as a unit normal.

Strictly speaking, the variance should be computed over all possible realizations of the data. This is not computationally feasible, but this "population" variance can be estimated by the "sample" variance within the data set. This estimate will be quite accurate when large numbers of sib pairs are used, which is the realistic setting even for the detection of moderate effects.

QTL Mapping: An Example

We have implemented generalized Haseman-Elston QTL regression analysis, ML QTL variance estimation, and nonparametric QTL analysis in MAPMAKER/SIBS. The use of these methods is illustrated in figure 8. The example shows the results of a simulation in which 1,000 sib pairs without available parents were genotyped for markers spaced every 10 cM on a 100-cM chromosome, each with five equally frequent alleles, corresponding to a heterozygosity of .8. A locus with environmental variance $\sigma_e^2 = 1.0$, additive genetic variance $\sigma_g^2 = 0.5$, and no dominance variance is located at 25 cM. Thus, 33% of the variance is explained by the QTL.

In figure 8*A* we compare the t statistics obtained by Haseman-Elston regression analysis by using regression on the distribution of v_i (*solid line*) and on the expected value $\hat{v}_i$ (*dashed line*). Note that the t statistic for the correct missing-value regression on the full distribution is always higher, indicating greater power. In figure 8*B* we compare the LOD scores obtained by ML variance estimation (*solid line*) and for Haseman-Elston regression on the distribution of v_i (*dashed line*). The LOD score is always higher for variance estimation, illustrating the greater power of the approach that uses the correct model. We would generally recommend the use of variance estimation over regression analysis. ML estimates of the three variances computed under the assumption of no dominance variance are plotted in figure 8C.

In figure 8*D* we plot Z scores produced by nonparametric mapping. The peak Z score is 2.66, reached at 25 cM (the true location of the locus). This Z score corresponds to a highly significant single-test p value of .004. However, it would not be statistically significant in a genomewide scan, since the threshold for a 5% false-positive rate is $Z = 4.1$, as described above. A Z score of ≥ 2.66 would be expected to occur by chance (i.e., in the absence of QTLs) about four times in the genome. This observation underscores the importance of using significance thresholds (single-point or genomewide) that are appropriate for the test being performed.

The example also emphasizes how difficult it is to prove linkage to a QTL in humans: even with a large effect (33% of variance explained) and a collection of 1,000 sib pairs, the statistical evidence by all three methods is still quite modest. (We assume an unselected sample of sibs. There has been discussion in the literature of the increased power of samples selected for extreme phenotypic differences [Carey and Williamson 1991; Cardon and Fulker 1994]. We do not address this issue here, other than to note that in applying the analytic methods described above, care must be taken to include the effects of selection. Simply using pairs with extreme phenotypic differences violates the underlying distributional assumption, greatly increases the risk of false-positive linkages, and inflates the population variance explained by the locus. One can solve the problem by incorporating the phenotypes from all pairs, even if the nonextreme pairs are not genotyped). The dramatic difference between QTL mapping in humans and experimental crosses among inbred strains is made evident by the fact that a similar QTL explaining 33% of the variance in a backcross involving 1,000 progeny would yield a LOD score of 55 (Lander and Botstein 1989).

Misspecification of Allele Frequencies

Marker allele frequencies do not enter into the calculation of IBD when parents are available for typing, but they matter when parents are unavailable. Misspecification of allele frequencies can produce false-positive results. Specifically, underestimating the frequency of an allele will lead to overestimating the degree of IBD sharing at the locus. Ideally, one should obtain good estimates of allele frequencies from the appropriate population. In addition, one should perform sensitivity analysis to test the sensitivity of results to changes in allele frequencies. To facilitate this process, MAPMAKER/SIBS

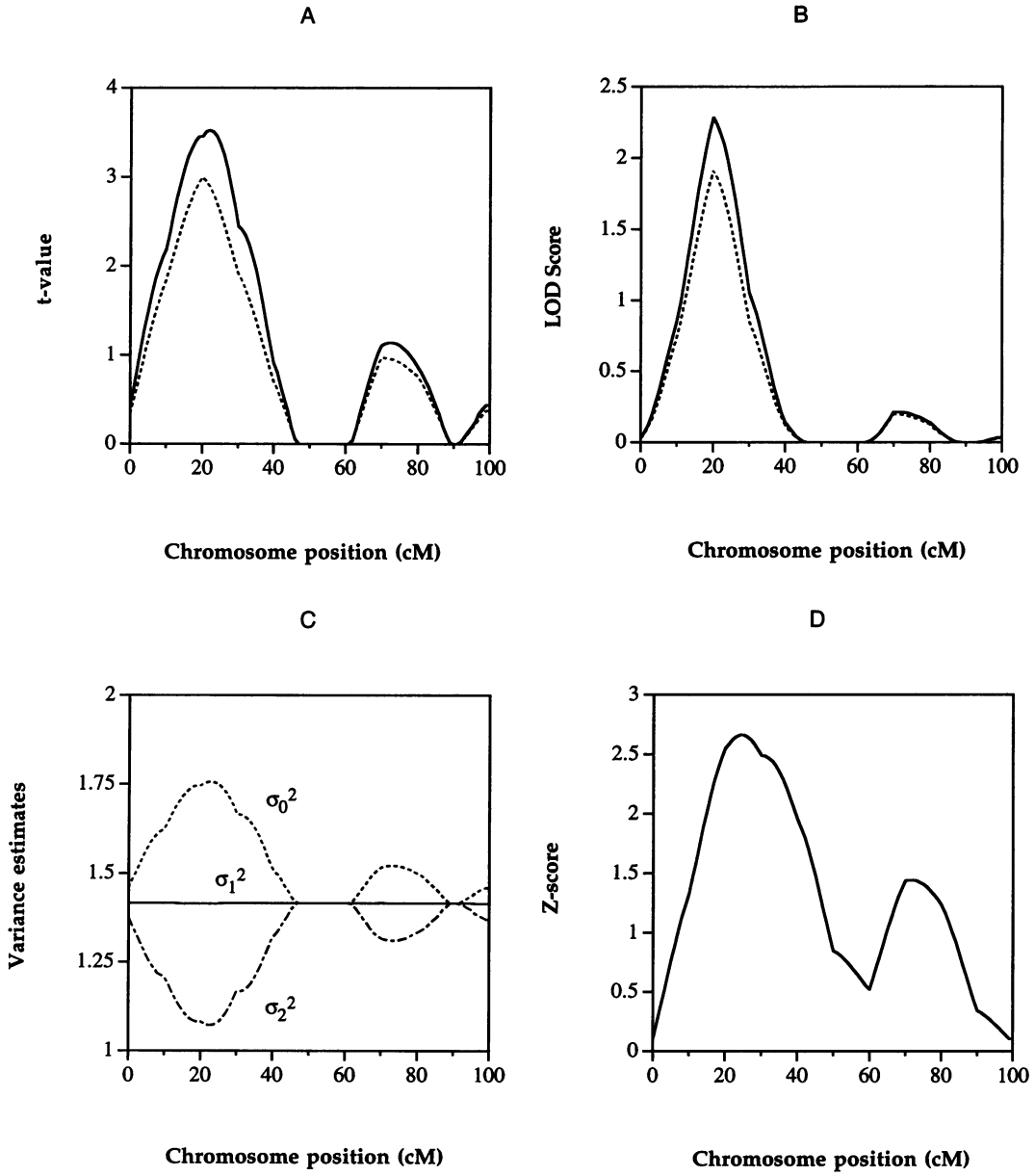

Figure 8 QTL mapping. Results of a simulation in which 1,000 sib pairs without available parents are genotyped for markers spaced every 10 cM on a 100 cM chromosome, each with five equally frequent alleles, corresponding to a heterozygosity of .8. A locus with $\sigma_e^2 = 1.0$ and $\sigma_g^2 = 0.5$ and no dominance variance is located at 25 cM. *A*, *t* statistics are plotted for Haseman-Elston regression analysis by using regression on the distribution of v_i (solid line) and on the expected value of v_i (*dashed line*). *B*, LOD scores are plotted for ML variance estimation (*solid line*) and for Haseman-Elston regression analysis by using regression on the distribution of v_i (*dashed line*). *C*, ML estimates of the variances computed under the assumption of no dominance variance. *D*, *Z* scores produced by nonparametric mapping

incorporates a feature allowing one to set a lower bound on the frequency of all alleles. Although the allele frequencies may sum to >1 in this sensitivity analysis, the corresponding LOD scores will be conservative.

Discussion

Sib-pair studies provide an increasingly important tool for genetic analysis of complex traits. Individual loci affecting a complex trait may have small effects and can be difficult to detect, requiring large studies. It is therefore desirable to have methods of analysis that make use of all available information. Current analysis methods study markers one at a time. In recent years, dense maps of genetic markers covering the entire human genome have become available (Buetow et al. 1994; Gyapay et al. 1994). Nonetheless, today's genetic markers are not infinitely polymorphic, and, thus, consider-

able additional power can be gained from studying multiple markers at once. This is especially true when parents are unavailable for typing, which is frequently the case, particularly with late-onset disorders.

We have now developed methods for carrying out complete multipoint sib-pair analysis. We use a rapid algorithm for multipoint likelihood computation, originally developed for homozygosity mapping (Kruglyak et al. 1995), to completely characterize the probability distribution of sharing 0, 1, or 2 alleles IBD for every sib pair at every point in the genome. This distribution contains all information available from marker genotypes. The knowledge of the IBD distribution allows every kind of statistical analysis to be carried out in a fully multipoint fashion.

In this paper, we develop methods for exclusion and ML mapping of qualitative traits. We also describe information-content mapping, a novel analytic technique that measures how much of the total inheritance information has been extracted by the markers and indicates where additional markers might be desirable. Using this technique, we examine the effects of map density, marker polymorphism, and availability of parents on the expected information content of a study. The results provide useful guidelines for study design. They also support the feasibility of a using a third-generation genetic linkage map based on plus-minus biallelic markers.

In addition, we develop a multipoint extension of the traditional QTL mapping approach of Haseman and Elston (1972), and describe two additional techniques for mapping quantitative trait loci: direct variance estimation by ML and nonparametric mapping. A number of other regression models have been proposed for QTL analysis (see, e.g., Cardon et al. 1994). All of these methods have performed regression on the expected number of alleles shared IBD, $\hat{v}_i$. As described above, the preferred procedure is to regress on the distribution of the number v_i of alleles shared IBD. Any regression model can be generalized in this fashion according to the EM procedure outlined in appendix B.

The analytical methods described above have been incorporated in a software package, MAPMAKER/SIBS. This package allows very rapid analysis and exploration of sib-pair data in a user-friendly environment. The MAPMAKER/SIBS package provides direct access to the full IBD distribution of v_i, making it straightforward to extend the package to any additional analysis techniques.

Previous approaches to sib-pair analysis have been partial solutions that fail to exploit the available data fully. The approach described here is the best possible, in the sense that it extracts *all* of the information about IBD status that can be gleaned from all markers and all individuals in the nuclear family. Inasmuch as the approach is also computationally rapid, there is no need to resort to partial solutions.

Finally, we note that these ideas and methods can be extended to general pedigree analysis. Such an extension will be described elsewhere.

Note added in proof.—An approximate multipoint approach to QTL mapping using regression on the expected number of alleles shared IBD was recently reported by Fulker et al. (1995).

Acknowledgments

We thank June Davies, John Todd, and colleagues for generously sharing data from their IDDM studies. We thank Mary Pat Reeve-Daly for programming assistance. This work was supported in part by National Institutes of Health grants HG00098 to E.S.L. and HG00017 to L.K.

Appendix A

EM Algorithm for Constrained Likelihood Maximization

We seek to maximize the likelihood

$$L = \prod_i \frac{z_0\rho_{i0} + z_1\rho_{i1} + z_2\rho_{i2}}{\alpha_0\rho_{i0} + \alpha_1\rho_{i1} + \alpha_2\rho_{i2}},$$

where ρ_{ij} is the probability of observing the data for the ith sib pair, given that the pair shares j alleles IBD, and z_j, the sharing proportions, are parameters to be estimated. We use the EM algorithm (Dempster et al. 1977; Little and Rubin 1987) to carry out the maximization and obtain ML estimates of the z_j (it is easily seen that the likelihood function belongs to an exponential-form family, and thus the EM algorithm applies).

In the E-step, we compute the expected sharing proportions in *each* pair according to

$$z_{ij} = \frac{z_i\rho_{ij}}{\sum_i z_i\rho_{ij}}.$$

We then sum the z_{ij} over the pairs to obtain the expected numbers n_0, n_1, n_2 of pedigrees sharing 0, 1, and 2 alleles IBD.

In the M-step, we use these expected numbers to obtain the new estimates of z_0, z_1, z_2. For unconstrained maximization, the estimates are trivial: $z_j = n_j/n$, where n is the total number of pairs.

The E- and M-steps are iterated until convergence to the (unconstrained) ML values of z_0, z_1, z_2. If the ML estimates fall within the possible triangle, they are accepted. If they fall outside the possible triangle, the likelihood is remaximized along one of the sides of the triangle (see Holmans [1993] for details). Here the E-step is

exactly as above. The M-step is slightly different: if the constraint is $z_1 = 1/2$, the new estimates are

$$z_0 = \frac{n_0}{2(n_0 + n_2)}; \quad z_2 = \frac{n_2}{2(n_0 + n_2)}.$$

If the constraint is $z_1 = 2z_0$, the new estimates are

$$z_0 = \frac{n_0 + n_1}{3n}; \quad z_1 = \frac{2(n_0 + n_1)}{3n}; \quad z_2 = \frac{n_2}{n}.$$

Constrained maximization can also be carried out from the start in order to test a particular model of inheritance.

Appendix B

Missing-Value Regression

In simple linear regression, one is given a set of N data points (x_i,y_i) assumed to be generated by $y = a + bx + \varepsilon$, where ε is a random normal variable with mean 0 and variance σ^2. Here, a, b, and σ^2 are unknown parameters to be estimated. The linear regression solutions are

$$a = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i}{N \Sigma x_i^2 - (\Sigma x_i)^2};$$

$$b = \frac{N \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{N \Sigma x_i^2 - (\Sigma x_i)^2};$$

and

$$\sigma^2 = \frac{1}{N} \Sigma (y_i - a - bx_i)^2.$$

These solutions are ML estimates, with the likelihood given by

$$L(a, b, \sigma^2) = \prod_i \phi(y_i - a - bx_i, \sigma^2),$$

where $\phi(t,\sigma^2)$ is the probability density at t of the normal distribution with mean 0 and variance σ^2.

In missing-value regression, the x_i's are not known with certainty. Instead, one has probability distributions over the possible values of x_i. The likelihood is then

$$L(a, b, \sigma^2) = \prod_i \sum_x p_i(x) L_i(x),$$

where i ranges over the data points, x ranges over the possible values of x_i, $p_i(x)$ denotes the probability that $x_i = x$, and $L_i(x)$ denotes the likelihood function for the ith data point, given that $x_i = x$. Standard linear regression cannot be performed. In particular, simply using the expected values $E[x_i]$ in place of the unknown x_i does *not* produce an ML solution. Instead, the likelihood function must be maximized explicitly.

This maximization can be easily accomplished by the EM algorithm (Dempster et al. 1977; Little and Rubin 1987), since the likelihood of the complete data belongs to an exponential-form family. In the E-step, one reestimates the probability distributions $p_i(x_i)$ *given* particular values of a, b, and σ^2 according to

$$p_i^{\text{est}}(x_i) = \frac{p_i(x_i)\phi(y_i - a - bx_i, \sigma^2)}{\sum_{\{x_i\}} p_i(x_i)\phi(y_i - a - bx_i, \sigma^2)},$$

and then computes the expected values of Σx_i, Σx_i^2, and $\Sigma x_i y_i$ conditional on the distributions. In the M-step, one substitutes the expected values of Σx_i, Σx_i^2, and $\Sigma x_i y_i$ in the simple regression formulas above to obtain new estimates of a, b, and σ^2. The E- and M-steps are iterated until convergence. The EM algorithm is guaranteed to converge to a local maximum of the likelihood function.

Appendix C

ML Estimation of Variance

When the number v_i of alleles shared IBD by each sib pair i is known, the likelihood function is

$$L = \prod_i \frac{1}{\sqrt{2\pi\sigma_{v_i}^2}} \exp\left(\frac{-D_i^2}{2\sigma_{v_i}^2}\right).$$

The ML estimates of the variances are easily seen to be

$$\sigma_j^2 = \frac{\sum_{\{i|v_i=j\}} D_i^2}{\sum_{\{i|v_i=j\}} 1},$$

i.e., the variance of D within each IBD class.

When the v_i's are not known with certainty, the likelihood function is

$$L = \prod_i \sum_j p_{ij} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-D_i^2}{2\sigma_j^2}\right),$$

where the sum is over possible allele sharing numbers j ($j = 0, 1, 2$), and p_{ij} is the probability that the ith pair shares j alleles IBD. This likelihood can be easily maximized by the EM algorithm (Dempster et al. 1977; Little and Rubin 1987) in a manner similar to that described

in appendix B. In brief, in the E-step, one re-estimates the probability distributions p_{ij} *given* the current estimates of σ_j^2 and then computes the expected values of $\Sigma_{\{i|v_i=j\}} D_i^2$ $(=\Sigma_i p_{ij} D_i^2)$ and $\Sigma_{\{i|v_i=j\}} 1$ $(=\Sigma_i p_{ij})$, conditional on the distributions. In the M-step, one substitutes these expected values in the simple formula for the variances above to obtain new estimates of σ_j^2. Constraints, such as the assumption of no dominance variance (i.e., $\sigma_1^2 = (\sigma_0^2 + \sigma_2^2)/2$), can be easily incorporated at this M-step. The E- and M-steps are iterated until convergence.

References

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

Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Heddema T, Duyk GM, Sheffield VC, Wang Z, et al (1994) Integrated human genome-wide maps constructed using the CEPH reference panel. Nat Genet 6:391–393

Cardon LR, Fulker DW (1994) The power of interval mapping quantitative trait loci, using selected sib pairs. Am J Hum Genet 55:825–833

Cardon LR, Smith SD, Fulker DW, Kimberling WJ, Pennington BF, DeFries JC (1994) Quantitative trait locus for reading disability on chromosome 6. Science 266:276–279

Carey G, Williamson J (1991) Linkage analysis of quantitative traits: increased power by using selected samples. Am J Hum Genet 49:786–796

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130–136

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B39:1–38

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, Cherry 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

Guo S-W (1994) Computation of identity-by-descent proportions shared by two siblings. Am J Hum Genet 54:1104–1109

Gyapay G, Morissette J, Vignal A, Dib C, Fizames C, Millasseau P, Marc S, et al (1994) The 1993–94 Généthon human genetic linkage map. Nat Genet 7:246–339

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

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

Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527

Kruglyak L, Lander ES (1995) A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421–1428

Lander ES, Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lander ES, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037–2048

Little RJA, Rubin DB (1987) Statistical analysis with missing data. Wiley, New York

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

Penrose LS (1938) Genetic linkage in graded human characters. Ann Eugenics 8:233–237

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

——— (1990*b*) 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

Sandkuijl LA (1989) Analysis of affected sib-pairs using information form extended families. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based upon affected pedigree members: genetic analysis workshop 6. Alan R Liss, New York

Terwilliger JD, Ott J (1994) Handbook of human genetic linkage. Johns Hopkins, Baltimore