Am. J. Hum. Genet. 58:1347–1363, 1996

Parametric and Nonparametric Linkage Analysis: A Unified Multipoint Approach

Leonid Kruglyak,[1] Mark J. Daly,[1] Mary Pat Reeve-Daly,[1] and Eric S. Lander[1,2]

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

Summary

In complex disease studies, it is crucial to perform multipoint linkage analysis with many markers and to use robust nonparametric methods that take account of all pedigree information. Currently available methods fall short in both regards. In this paper, we describe how to extract complete multipoint inheritance information from general pedigrees of moderate size. This information is captured in the multipoint inheritance distribution, which provides a framework for a unified approach to both parametric and nonparametric methods of linkage analysis. Specifically, the approach includes the following: (1) Rapid exact computation of multipoint LOD scores involving dozens of highly polymorphic markers, even in the presence of loops and missing data. (2) Nonparametric linkage (NPL) analysis, a powerful new approach to pedigree analysis. We show that NPL is robust to uncertainty about mode of inheritance, is much more powerful than commonly used nonparametric methods, and loses little power relative to parametric linkage analysis. NPL thus appears to be the method of choice for pedigree studies of complex traits. (3) Information-content mapping, which measures the fraction of the total inheritance information extracted by the available marker data and points out the regions in which typing additional markers is most useful. (4) Maximum-likelihood reconstruction of many-marker haplotypes, even in pedigrees with missing data. We have implemented NPL analysis, LOD-score computation, information-content mapping, and haplotype reconstruction in a new computer package, GENEHUNTER. The package allows efficient multipoint analysis of pedigree data to be performed rapidly in a single user-friendly environment.

Received January 12, 1996; accepted for publication March 4, 1996.

Address for correspondence and reprints: Dr. Eric S. Lander or Dr. Leonid Kruglyak, Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142. E-mail: lander@genome.wi.mit.edu or leonid@genome.wi.mit.edu

0002-9297/96/5806-0028$02.00

Introduction

Linkage analysis aims to extract all available inheritance information from pedigrees and to test for coinheritance of chromosomal regions with a trait. In principle, one can use either parametric methods, which involve testing whether the inheritance pattern fits a specific model for a trait-causing gene, or nonparametric methods, which involve testing whether the inheritance pattern deviates from expectation under independent assortment.

Although easily stated, this goal has proved hard to implement in practice. A major obstacle has been the computational difficulty of making inferences based on imperfect information, arising from incomplete structure of human pedigrees and incomplete informativeness of genetic markers. Parametric and nonparametric methods have generally adopted rather different solutions, neither of which is wholly satisfactory:

1. *Parametric analysis.* The LOD-score method is the most widely used approach to parametric linkage analysis (Morton 1955); its theoretical foundations are well understood, and computer programs to carry out LOD-score calculations are available (Ott 1991; Terwilliger and Ott 1994). The major difficulty is computational—extracting the full linkage information in a pedigree requires the use of a dense genetic linkage map, but such multipoint analysis is infeasible for more than a handful of loci because of the inherent constraints of the Elston-Stewart algorithm (Elston and Stewart 1971). The problem has been circumvented in the case of specific pedigree structures, through the use of alternative algorithms (Lathrop et al. 1986; Lander and Green 1987; Kruglyak et al. 1995), and recent improvements to the Elston-Stewart algorithm promise to make multipoint analysis with a limited number of loci more practical (O'Connell and Weeks 1995). Nonetheless, complete multipoint analysis remains a bottleneck for general pedigrees—even those of moderate size.
2. *Nonparametric analysis.* Because parametric linkage analysis can be highly sensitive to misspecification of the linkage model (Clerget-Darpoux et al. 1986), nonparametric analysis is a key tool for all but the simplest of traits. Nonparametric analysis has been performed primarily by one of two methods. The

first approach is to break pedigrees into nuclear families and apply sib-pair analysis; this is inefficient because it wastes a great deal of inheritance information contained in pedigree structure. To partly utilize pedigree information, Weeks and Lange (1988, 1992) developed the affected-pedigree-member method (APM). APM is not a true linkage method. It sidesteps the thorny issue of tracing the inheritance pattern in a pedigree by focusing on whether affected relatives happen to show the same alleles at a locus (i.e., identity/identical by state [IBS]), regardless of whether the allele is actually inherited from a common ancestor (i.e., identity/identical by descent [IBD]). The extent of IBS sharing among all pairs of affected members of the pedigree is compared with Mendelian expectation under the hypothesis of no linkage. The APM approach has several drawbacks: (i) It focuses only on IBS information and ignores genotype information for additional members of the pedigree, even when this information can be used to resolve whether shared alleles are actually IBD. (ii) It involves comparisons only among pairs of individuals, which can be less powerful than tests based on larger sets of affected individuals (Whittemore and Halpern 1994*a*; also, see below). (iii) It lacks a true multipoint formulation. Multilocus APM simply adds together statistics from several marker loci (Weeks and Lange 1992), rather than extracting linkage information about any given point. It thus tests for linkage to an extended chromosomal region rather than to a point, and therefore it cannot be used to localize a particular locus relative to a map markers. By failing to extract the full inheritance information, APM is potentially prone to false-positive and false-negative results.

To avoid these inherent problems of IBS-based methods, Curtis and Sham (1994) have recently proposed an approach, called *extended relative pair analysis* (ERPA), that uses the risk-calculation facility of the LINKAGE package (Lathrop et al. 1984) to compute IBD-sharing probabilities for all pairs of affected individuals in a pedigree. ERPA is thus a true linkage approach to nonparametric analysis. It is limited, however, in several key respects: the comparisons are inherently confined to relative pairs; the statistical test for linkage is ad hoc; and the method cannot handle large numbers of loci, because of the basic algorithm used in the LINKAGE package. Other approaches to nonparametric analysis have also been described (e.g., by Curtis and Sham 1995).

The purpose of this paper is to describe a unified approach to both parametric analysis and nonparametric analysis. The key is to separate two issues: (1) extracting information about the inheritance pattern in a pedigree (which depends only on the genetic markers) and (2) defining a statistic to assess linkage for a given inheritance pattern (which depends only on the nature of the trait).

This approach generalizes our recent methods for complete multipoint sib-pair analysis (Kruglyak and Lander 1995) to the situation of arbitrary pedigrees. The generalization required the development of a new linkage algorithm for arbitrary pedigrees, as well as the definition of new statistics for performing nonparametric analysis.

The paper is organized in four parts. First, we discuss how to extract all available inheritance information from a pedigree. Specifically, we present a complete multipoint algorithm for determining the probability distribution over possible inheritance patterns at each point in the genome. Second, we apply these concepts to define a unified multipoint framework for both parametric and nonparametric analysis. In the former case, the approach provides a rapid multipoint linkage algorithm for traditional LOD-score calculations. In the latter case, it provides a powerful new approach to pedigree analysis, which we refer to as *nonparametric linkage* (NPL) *analysis*. Third, we evaluate the power of NPL analysis in applications to both simulated and actual data. In all cases examined, NPL analysis is considerably more powerful than APM. Finally, we show how the framework presented here also allows reconstruction of haplotypes in pedigrees.

We have implemented these methods in a computer program, GENEHUNTER, for both parametric and nonparametric analysis. With current workstations, the program can rapidly analyze moderately sized pedigrees of the sort used in genetic studies of complex traits.

Definitions

Given a pedigree, we define *nonfounders* to be those individuals whose parents are in the pedigree. Without loss of generality, we will assume that pedigrees are defined to include both parents of any individual who has a sib, half-sib, or parent in the pedigree. (If such parents are unavailable for study, they are simply included in the pedigree with unknown phenotypic and genotypic status). Individuals whose parents are not in the pedigree are designated as *founders*. Throughout, n will denote the number of nonfounders, and f the number of founders, in a pedigree. Founders will be assumed to be unrelated; that is, they are assumed to carry $2f$ alleles that are distinct by descent (although some may be IBS).

Representing and Computing Inheritance Information

The Inheritance Vector

Linkage analysis can be divided into two steps: (i) inferring information about the inheritance pattern of a

pedigree and (ii) deciding whether the inheritance information indicates the presence of a trait-causing gene. Ideally, one would like to know the precise inheritance pattern at every locus in the genome. The inheritance pattern at each point x is completely described by a binary *inheritance vector* $v(x) = (p_1,m_1,p_2,m_2, \ldots,p_n,m_n)$, whose coordinates describe the outcome of the paternal and maternal meioses giving rise to the n nonfounders in the pedigree (Lander and Green 1987). Specifically, $p_i = 0$ or 1, according to whether the grandpaternal or grandmaternal allele was transmitted in the paternal meiosis giving rise to the ith nonfounder; m_i carries the same information for the corresponding maternal meiosis. Thus, the inheritance vector completely specifies which of the $2f$ distinct founder alleles are inherited by each nonfounder. The notion of the inheritance vector is illustrated in figure 1*A*. The set of all 2^{2n} possible inheritance vectors will be denoted V. Similar representations of inheritance have been proposed in the context of Monte Carlo linkage analysis (Sobel and Lange 1993; Thompson 1994), as well as in other applications (Whittemore and Halpern 1994*a*, 1994*b*; Guo 1995).

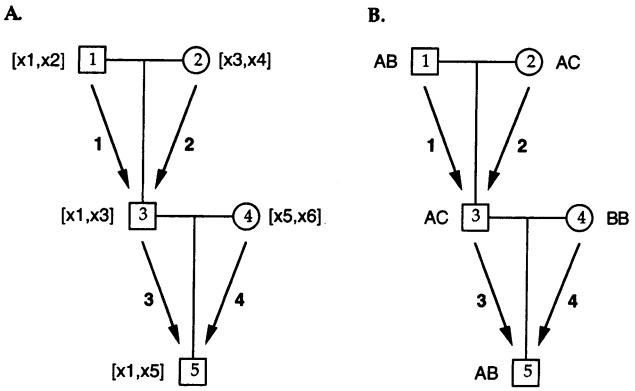
The Inheritance Distribution

In practice, it is not feasible to determine the true inheritance vector at every point in the genome, since this would require genotyping all pedigree members with an infinitely dense map of fully informative markers. Because key pedigree members are frequently unavailable and genetic markers have limited heterozygosity, genotype data will provide only partial information about inheritance.

Partial information extracted from a pedigree can be represented by a probability distribution over the possible inheritance vectors at each locus in the genome—that is, $P(v(x) = w)$ for all inheritance vectors $w \in V$. In the absence of any genotype information, all inheritance vectors are equally likely according to Mendel's first law, and the probability distribution is uniform (abbreviated as P_{uniform}). As genotype information is added, the probability distribution is concentrated on certain inheritance vectors. The probability distribution over possible inheritance vectors will be referred to as the *inheritance distribution;* the notion is illustrated in figure 1*B* and *C*.

Calculating the Inheritance Distribution by Use of Hidden Markov Models (HMMs)

To extract the full information from a data set, one must calculate the inheritance distribution conditional on the genotypes at all marker loci (abbreviated P_{complete}). Lander and Green (1987) described how, in principle, an HMM can be used to solve this problem. In brief, the approach considers the inheritance pattern across the genome as a Markov process (with recombination causing transitions among states) that is observed, imperfectly, only at marker loci. One uses the imperfect observations at each marker (more precisely, the probability distribution over inheritance vectors at each marker locus, conditional only on the data for the locus itself [abbreviated as P_{marker}]), to reconstruct the probability distribution at any point, conditional on the entire data set, according to the standard forward-backward conditioning approach employed in HMMs (Rabiner 1989). In the basic Lander-Green algorithm, the time



C.

inheritance vector	prior	posterior	true
0000	1/16	1/8	1
0001	1/16	1/8	0
0010	1/16	0	0
0011	1/16	0	0
0100	1/16	1/8	0
0101	1/16	1/8	0
0110	1/16	0	0
0111	1/16	0	0
1000	1/16	1/8	0
1001	1/16	1/8	0
1010	1/16	0	0
1011	1/16	0	0
1100	1/16	1/8	0
1101	1/16	1/8	0
1110	1/16	0	0
1111	1/16	0	0

Figure 1 Illustration of the inheritance vector and its distribution, for a simple pedigree. *A*, Pedigree shown with individuals labeled "1" through "5." The distinct-by-descent founder alleles are labeled "x1" through "x6"; they are assumed to be phase known, with the paternally derived allele listed first. The four meiotic events whose outcomes determine inheritance in the pedigree are indicated by arrows; the labels correspond to the coordinates in the inheritance vector. The inheritance outcome shown is specified by inheritance vector (0,0,0,0)—that is, the paternally derived allele is transmitted in every meiosis. *B*, Same pedigree, now shown with actual genotypes at a marker with three alleles, A, B, and C. Only the outcome of meiosis 3 is unambiguously determined by the genotype data—the paternally derived allele is transmitted, fixing the third bit in the inheritance vector at 0. *C*, Inheritance distribution for the 16 possible inheritance vectors. "prior" denotes distribution before any genotyping has been performed; "posterior" denotes distribution based on genotypes in panel *B;* and "true" denotes distribution based on fully informative, phase-known genotypes as in panel *A*.

required for the HMM reconstruction step with m markers is $O(m \cdot 2^{4n})$. Because this scales linearly with the number of loci but exponentially with the number of nonfounders, the approach is best suited to complete multipoint analyses in pedigrees of moderate size. In contrast, the Elston-Stewart algorithm scales exponentially with loci but linearly with nonfounders and thus is best suited for studying one or a few markers in large pedigrees.

First Speedup

Kruglyak et al. (1995) recently showed how to decrease the time required for the HMM reconstruction step from $O(m \cdot 2^{4n})$ to $O(m \cdot n2^{2n})$, thereby effectively doubling the pedigree size to which the HMM approach can be applied. With this speedup, the approach has been implemented in special cases, to allow complete multipoint analysis for homozygosity mapping, linkage analysis in nuclear families, and sib-pair analysis (Kruglyak et al. 1995; Kruglyak and Lander 1995). To apply the approach to general pedigrees, it is necessary to have an algorithm for calculating the initial distributions used in the HMM, P_{marker}, for pedigrees of arbitrary structure. We have now devised such an algorithm, which is described in appendix A.

Second Speedup

We have devised a further substantial acceleration of the HMM, by taking advantage of a certain degeneracy among the inheritance vectors. Since a pedigree contains no information about founder phase, inheritance vectors that differ only by phase changes in the founders are completely equivalent and must therefore have equal probabilities. In a pedigree with f founders, the inheritance vectors can thus be organized into equivalence classes consisting of 2^f equivalent members. The HMM algorithm can be modified to work with just a single representative from each equivalence class, as described in appendix B. This reduces both the time and space requirements of the calculation by a factor of 2^f, further increasing the size of pedigrees that may be analyzed. The running time for analysis of m markers is thus $O(m \cdot n2^{2n-f})$.

Computer Implementation

We have implemented the HMM approach with these two speedups in a new computer package, GENEHUNTER. On current workstations, GENEHUNTER can comfortably handle pedigrees with $2n - f \leq 16$, or, typically, approximately a dozen nonfounders. Some examples of pedigrees that can be readily analyzed are given in figure 2.

The same methods also can be used to estimate the number of recombination events between two markers. GENEHUNTER includes an option to compute this number for pairs of consecutive markers, which can be useful for detecting genotyping errors that cause map inflation.

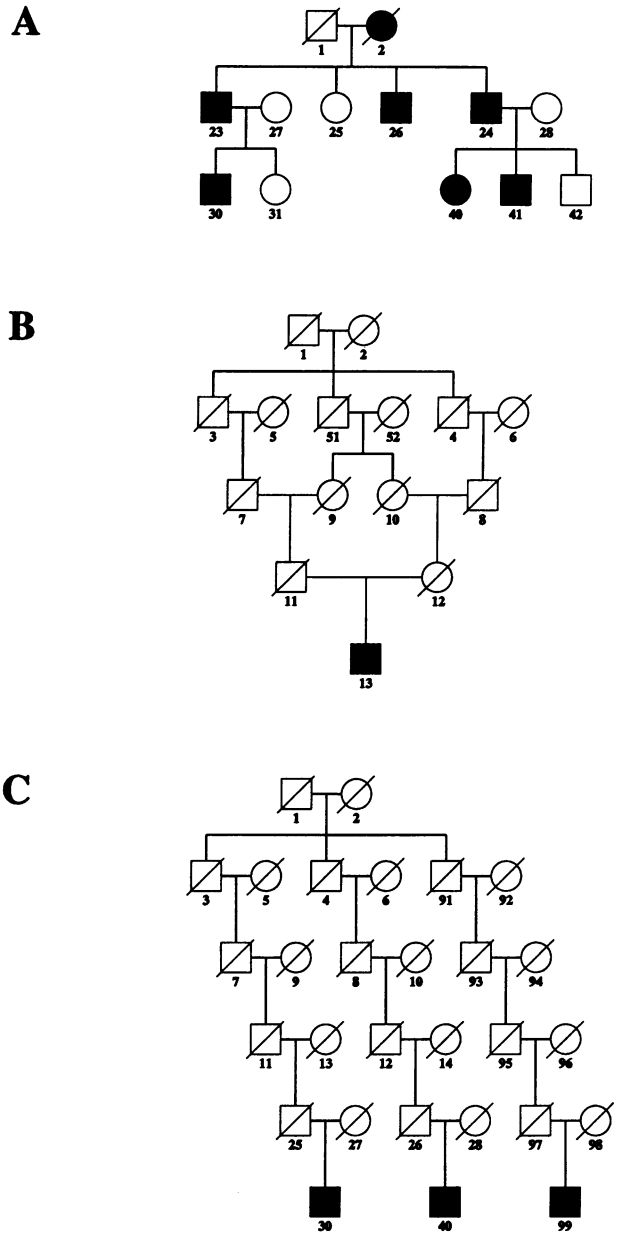


Figure 2 Examples of pedigrees that can be analyzed by using GENEHUNTER. *A*, Simple three-generation pedigree segregating a dominant disorder. Individuals in the last two generations are available for study. *B*, Complex inbred pedigree that occurred in the study of Werner syndrome (Thompson and Wijsman 1994). Only the affected individual in the last generation is available for study. *C*, Pedigree with three affected fourth cousins. Only the affected individuals in the last generation are available for study.

Information-Content Mapping

In studying a pedigree, it is useful to know how much of the total inheritance information has been extracted

at each point in the genome, given the available genotype data. We introduced a notion of "information-content mapping" in our previous work on sib-pair analysis (Kruglyak and Lander 1995). Information content provides a measure of how closely a study approaches the goal of completely determining the inheritance outcome, and it points out the regions where typing additional markers is most useful. Here, we modify our previous approach and extend it to arbitrary pedigrees.

The classical information-theoretic measure of residual uncertainty in a probability distribution is its entropy, defined by $E = -\Sigma P_i \log_2 P_i$, where P_i is the probability of the ith outcome and where $\log_2$ is used in order for the entropy to be measured in bits (Shannon 1948). The entropy of the probability distribution over inheritance vectors thus naturally reflects information content.

In the absence of genotype data, the probability distribution is uniform over all 2^{2n-f} equivalence classes of inheritance vectors. The entropy of the distribution is easily seen to be $E = 2n - f$ bits. This result makes intuitive sense, since we are completely uncertain about the outcome of the $2n - f$ meioses for which information can be obtained. If the inheritance vector is known with certainty (e.g., at a fully informative marker), the probability distribution is completely concentrated on a single outcome. The entropy is thus $E = 0$, which again makes intuitive sense.

The information content of the inheritance pattern at point x will be defined by

$$I_E(x) = 1 - E(x)/E_0 , \quad (1)$$

where $E(x)$ is the entropy of the multipoint inheritance distribution at x and where $E_0 = 2n - f$ bits is the entropy in the absence of genotype data. Information content $I_E(x) = 1$ indicates perfect informativeness at x, whereas information content $I_E(x) = 0$ indicates total uncertainty about inheritance in the pedigree at x. Since entropy is an additive measure, it can be summed over all pedigrees in the data set. Equation (1) is then used with total entropy to obtain the overall information content of a study.

I_E is a general measure of information content. It does not depend on any particular test for linkage and has the desirable property that it always lies between 0 and 1. (This contrasts with a somewhat different measure of information content, which we discussed in previous work on sib-pair analysis [Kruglyak and Lander 1995].) An example of information content for different map densities is shown in figure 3.

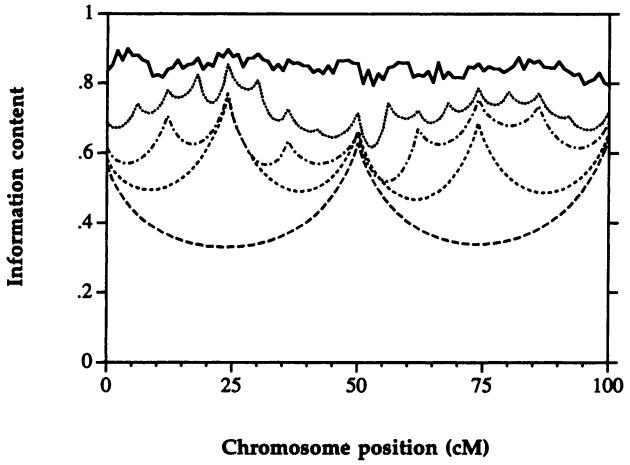


Figure 3 Information-content mapping for various marker densities. Genotypes for markers spaced every 2 cM on a 100-cM map, with typical microsatellite informativeness levels (heterogeneity ~.75), were simulated for 10 sibships having four sibs each, with missing parents. The five curves show the information content with markers genotyped at 50-cM, 25-cM, 12-cM, 6-cM, and 2-cM average spacing (corresponding, respectively, to 3, 5, 9, 17, and 51 markers genotyped across the map). The average information content increases from ~40% to 54%, 63%, 72%, and 85%, respectively.

Unified Linkage Analysis

We now define both parametric and nonparametric analysis from a unified perspective, which is based on the notion of inheritance vectors. In the ideal situation—the precise inheritance vector $v(x)$ at each point x is known with certainty—linkage analysis simply involves quantifying the extent to which the inheritance vector indicates the presence of a disease gene. This can be done by specifying a scoring function $S(v,\Phi)$ that depends on the inheritance vector v and the observed phenotypes Φ in the pedigree.

To extend the analysis to the more realistic situation in which one has only a probability distribution over $v(x)$, one can generalize the scoring function by taking its expected value over the inheritance distribution:

$$\bar{S}(x,\Phi) = \sum_{w \in V} S(w,\Phi) P[v(x) = w] . \quad (2)$$

Given the probability distribution over inheritance vectors at every point x, it is then straightforward to calculate $\bar{S}$ throughout the genome. Specifically, one could calculate once and store the 2^{2n-f} values of $S(v,\Phi)$. For each point x, one could then compute the linear combination in equation (2) in time $O(2^{2n-f})$. We now consider various choices of scoring functions S that correspond to parametric linkage analysis and NPL analysis.

Parametric Linkage Analysis

Scoring Function

In parametric linkage analysis, one assumes a model describing the probability of phenotype given genotype

at the disease locus and calculates the likelihood ratio under the hypothesis that a disease gene is at x, versus the hypothesis that it is unlinked to x. In the special case when the inheritance vector is known, the scoring function S is simply the likelihood ratio. It is given by

$$\mathrm{LR}(v) = \frac{P(\Phi \mid v)}{\sum_{w \in V} P(\Phi \mid w) P_{\mathrm{uniform}}(w)}.$$

$P(\Phi \mid v)$ is simply the likelihood of observed phenotypes Φ, conditioned on the particular inheritance vector v; it depends only on the penetrance values and allele frequencies at the disease locus. For each v, one can efficiently compute $P(\Phi \mid v)$ by a simple adaptation of standard peeling methods for pedigrees without loops (Elston and Stewart 1971; Lange and Elston 1975; Cannings et al. 1978; Whittemore and Halpern 1994*b*) and by a combination of peeling, loop breaking, and enumeration of founder genotypes for pedigrees with loops (for details, see appendix C). Calculating the likelihood for each of the 2^{2n-f} equivalence classes of inheritance vectors is rapid for moderate-sized pedigrees, both with and without loops.

In the general case, we take the expectation of the scoring function over the inheritance distribution, as in equation (2):

$$\begin{aligned}\overline{\mathrm{LR}}(x) &= \sum_{w \in V} \mathrm{LR}(w) P(v(x) = w) \\ &= \frac{\sum_{w \in V} P(\Phi \mid w) P_{\mathrm{complete}}(w)}{\sum_{w \in V} P(\Phi \mid w) P_{\mathrm{uniform}}(w)}.\end{aligned}$$

This expression is easily seen to be equivalent to the traditional definition of the likelihood ratio—the numerator is proportional to the multipoint likelihood when the disease gene is at x, whereas the denominator is proportional to the unlinked likelihood. According to long-standing tradition, one reports the LOD score, $\log_{10}(\overline{\mathrm{LR}})$.

Because traditional LOD-score analysis can be expressed in the unified framework above, the fast HMM approach provides a rapid algorithm for performing complete multipoint linkage analysis in moderate-sized pedigrees. The LOD scores obtained by this method are exact—no approximations are involved. The only difference with conventional algorithms is the speed of computation when many markers are considered simultaneously.

Implementation

We have implemented parametric linkage analysis within GENEHUNTER. The program can compute LOD scores for arbitrary pedigrees under particular models of inheritance, allowing the user to specify allele frequencies at the disease locus and penetrances for liability classes (including age- and sex-dependent penetrances). The program also allows the user to test for linkage under genetic heterogeneity by using an admixture model (Ott 1991; Terwilliger and Ott 1994) to estimate the proportion of linked families α. Alternatively, the user can specify the admixture parameter α.

To illustrate its performance, GENEHUNTER was applied to simulated data for the pedigrees shown in figure 2. For each pedigree, we simulated genotype data for a genetic map of 20 markers under the hypothesis of a disease-causing gene located in the middle of the map. We then calculated complete multipoint LOD scores at each marker and at four points within each interval between markers, that is, at 96 distinct map locations (fig. 4). On a DEC Alpha workstation, the computation times for these 96 21-point LOD scores (disease locus plus all 20 markers) were 24 min, 82 min, and 280 min, for pedigrees A, B, and C, respectively. (The respective values of $2n - f$ are 14, 15, and 16).

For each of the three pedigrees, the maximum LOD score computed by using complete multipoint analysis approaches the theoretical maximum LOD score that would be obtained with an infinitely polymorphic marker located at a recombination fraction of zero from the disease gene. In particular, for pedigree C in figure 2, the three isolated fourth cousins have a probability of $(\frac{1}{2})^{13}$ of sharing an allele IBD, resulting in a theoretical maximum LOD score of 3.91. The multipoint LOD score nearly achieves this maximum, with a LOD of 3.84 (fig. 4C), indicating that it has extracted essentially all inheritance information. In contrast, the maximum LOD score attainable with a single marker is only 1.87, and the maximum LOD score with two flanking markers is 1.98. In this case, multipoint analysis increases the LOD score from moderately interesting to significant, providing almost 100-fold-higher odds in favor of linkage than does two-point analysis.

To further explore the value of multipoint analysis, we considered the simpler case of a pedigree with two affected fourth cousins and all other pedigree members unavailable for study. We once again simulated a 20-marker map under the hypothesis of a linked rare dominant gene. The IBD-sharing probability for two fourth cousins is 1/256, yielding a theoretical maximum LOD score of 2.41. In figure 5, we plot the maximum LOD score achieved by analyzing $k = 1, \ldots, 20$ consecutive markers simultaneously. Complete 20-marker analysis yields a LOD score of 2.2 (91% of theoretical maximum). In contrast, the highest two-point LOD score is only 0.83 (34% of theoretical maximum), and even simultaneous six-marker analysis yields, at most, a LOD score of 1.74 (72% of theoretical maximum). These results underscore the value that multipoint analysis with

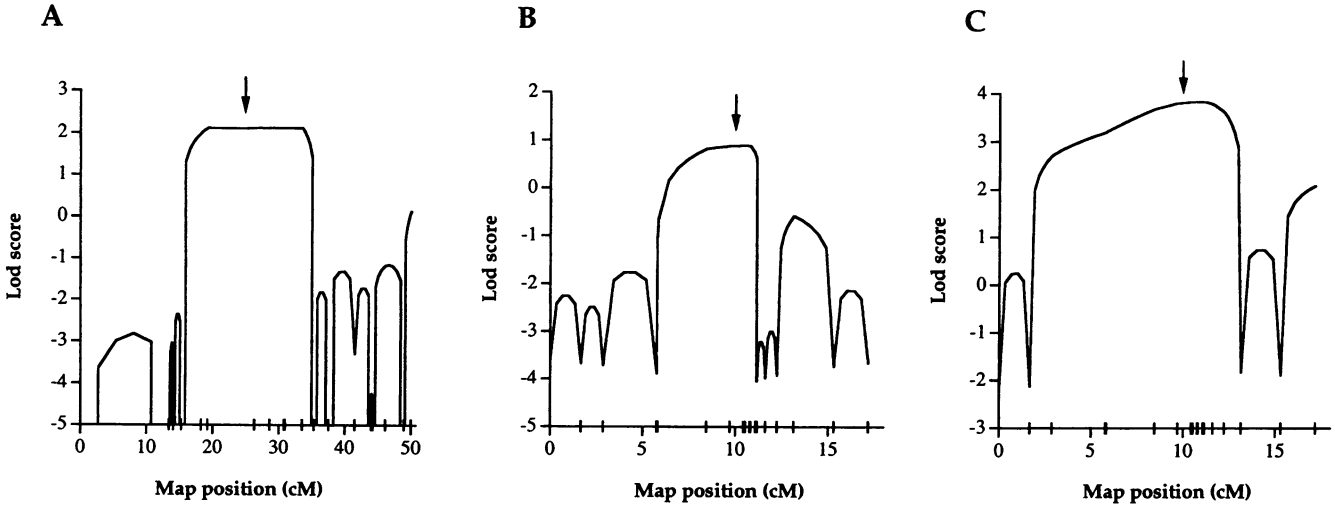


Figure 4 Multipoint LOD-score plots for the pedigrees shown in figure 2. Genotypes for 20 markers were simulated under the assumption of a disease gene at the location indicated by an arrow. A total of 96 21-point LOD scores were computed, with the disease locus tested at each marker and at four evenly spaced locations in each interval between markers. Marker positions are indicated by tick marks on the horizontal axis. *A*, Pedigree of figure 2*A*, with a rare dominant gene (frequency 10^{-4}). *B*, Pedigree of figure 2*B*, with a rare recessive gene (frequency 10^{-4}). *C*, Pedigree of figure 2*C*, with a very rare dominant gene (frequency 10^{-6}).

many markers has for extracting the full inheritance information. Such multipoint analysis is clearly desirable, since it requires only 40 s on a SUN SPARC workstation running GENEHUNTER.

To compare the performance of GENEHUNTER with that of other linkage packages, we analyzed the pedigree with two affected fourth cousins, using FASTLINK (Cottingham et al. 1993) and VITESSE (O'Connell and Weeks 1995), both running on a SUN SPARC workstation. FASTLINK required 32 min to compute LOD scores when using overlapping sets of two markers (28 three-point calculations), with a maximum LOD score of 0.98. Four-point calculations failed to complete after ~100 h. VITESSE required 85 s to compute LOD scores when using two markers simultaneously, 30 min to compute LOD scores when using three markers simultaneously (54 four-point calculations; maximum LOD score of 1.28), and 19 h 14 min to compute lod scores when using four markers simultaneously (68 five-point calculations; maximum LOD score of 1.43). Six-point calculations failed to complete after ~100 h. These other programs thus can perform multipoint analysis with a handful of markers, but not the complete multipoint calculations necessary to extract all available inheritance information. On the other hand, these programs are able to handle very large pedigrees that are beyond the computational limitations of GENEHUNTER.

GENEHUNTER's speed is independent of the number of alleles per marker (thereby allowing highly polymorphic markers to be used without recoding) and is essentially independent of the amount of missing information in the pedigree. The program has been tested extensively by comparing the results with those produced by LINKAGE (Lathrop et al. 1984) and FASTLINK (Cottingham et al. 1993), for a variety of family structures and modes of inheritance (in analyses using a small number of markers). In all case examined, the three programs produced identical answers.

NPL Analysis

Scoring Functions

We begin by considering the special case in which the inheritance vector is known with certainty. The inheritance vector fully determines which of the $2f$ distinct founder alleles was inherited by each person and thus completely specifies IBD sharing in the pedigree. The only issue is to define a suitable scoring function to measure whether affected individuals share alleles IBD more often than expected under random segregation. One simple approach would be to assign a score of 1 if all affected individuals in a pedigree share an allele IBD and to assign a score of 0 otherwise (Thomas et al. 1994). However, this statistic is likely not to be robust in the presence of phenocopies and common disease alleles. We consider below two useful scoring functions, S_{pairs} and S_{all}, previously discussed by Whittemore and Halpern (1994*a*); other scoring functions can be defined.

1. IBD sharing in pairs.—One possible approach is to count pairwise allele sharing among affected relatives. Given the inheritance vector v, $S_{\text{pairs}}(v)$ is defined to be the number of pairs of alleles from distinct affected pedigree members that are IBD. The traditional APM statistic (Weeks and Lange 1988) also counts pairwise allele

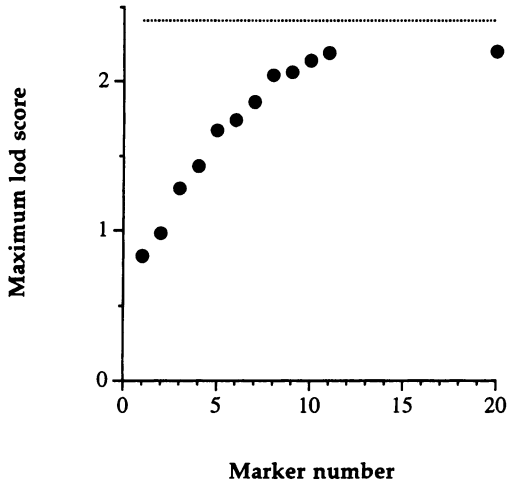


Figure 5 Maximum LOD score achieved in a pedigree with two affected fourth cousins, plotted as a function of k, the number of markers analyzed simultaneously. Genotypes for 20 markers were simulated by assuming the presence of a very rare dominant disease gene (frequency 10^{-6}) in the middle of an 18-cM map, as in figure 4C. LOD scores were computed with the disease locus tested at the markers and at four points within each interval between markers; genotype data from overlapping sets of k consecutive markers were used. Black dots show results for multipoint analysis with sets of 1,2,3, . . . ,11, and 20 markers; and the dotted line shows the theoretical maximum LOD score of 2.41 for this pedigree. The maximum LOD score of 2.2, achieved with 20 markers, is the highest possible with this marker density and polymorphism; doubling the marker density and performing multipoint analysis with 40 markers raises the maximum LOD score to 2.4 (data not shown).

sharing, but it is based on sharing IBS rather than on sharing IBD; the two statistics will coincide only at markers for which IBS unambiguously determines IBD.

2. IBD sharing in larger sets.—One can often increase statistical power by considering larger sets of affected relatives, rather than just pairs. For example, it is more impressive to find that five affected relatives share the *same* allele IBD than to find that each pair of them shares *some* allele IBD. Whittemore and Halpern (1994*a*) proposed an interesting statistic to capture the allele sharing associated with a given inheritance vector v. Let a denote the number of affected individuals in the pedigree, let h be a collection of alleles obtained by choosing one allele from each of these affected individuals, and let $b_i(h)$ denote the number of times that the ith founder allele appears in h (for $i = 1, \ldots, 2f$). The score S_{all} is defined as

$$S_{\text{all}}(v) = 2^{-a} \sum_{h} \left[\prod_{i=1}^{2f} b_i(h)! \right],$$

where the sum is taken over the 2^a possible ways to choose h. In effect, the score is the average number of permutations that preserve a collection obtained by choosing one allele from each affected person. It gives sharply increasing weight as the number of affected individuals sharing a particular allele increases.

For either approach, we define a normalized score

$$Z(v) = [S(v) - \mu]/\sigma, \quad (3)$$

where μ and σ are the mean and SD of S under P_{uniform}, the uniform distribution over the possible inheritance vectors. (These quantities can be calculated by enumeration over all vectors.) Under the null hypothesis of no linkage (i.e., P_{uniform}), the normalized score Z has mean 0 and variance 1.

To combine scores among m pedigrees, one can take a linear combination

$$Z = \sum_{i=1}^{m} \gamma_i Z_i, \quad (4)$$

where m is the number of pedigrees, Z_i denotes the normalized score for the ith pedigree, and the γ_i are weighting factors. The weighting factors should be chosen so that $\Sigma_i \gamma_i^2 = 1$, so that Z has mean 0 and variance 1 under the null hypothesis of no linkage. We will use $\gamma_i = 1/\sqrt{m}$ in the applications below; this choice appears to provide a good compromise between small and large pedigrees. It may be possible to increase power by selecting γ_i according to the nature of the pedigrees, but we will not explore this issue here, other than to note that the optimal choice will likely depend on the (usually unknown) genetic architecture of particular diseases.

We will refer to Z as the *NPL score* for the collection of pedigrees. In some cases, we will speak of $\text{NPL}_{\text{pairs}}$ and NPL_{all} scores, to indicate the scoring function under consideration.

Statistical Significance

Suppose that analysis of pedigrees yields an NPL statistic of Z_{obs}. What is the significance level of this observation? There are two simple approaches:

1. *Exact distribution.* It is straightforward to compute the exact probability distribution of the overall score Z under the null hypothesis of no linkage. Specifically, one can calculate the distribution for each pedigree by enumerating all possible inheritance vectors; the distribution for the collection of pedigrees is then obtained by convolving these distributions. One can then simply look up the exact value, $P(Z \geq Z_{\text{obs}})$.
2. *Normal approximation.* Under the null hypothesis of no linkage, the score Z will tend toward a standard normal variable as one studies many similar pedigrees. (This follows from the central limit theorem, since Z is an appropriately normalized sum of inde-

pendent random variables.) The significance level of an observation Z_{obs} can then be approximated by consulting a table of tail probabilities for the standard normal. Although less precise than the exact distribution, the normal approximation is useful in some settings.

Imperfect Data

We have so far considered the situation in which the inheritance vector is known with certainty. In fact, it is straightforward to extend Z to the general case, by taking its expected value over the inheritance distribution, as in equation (2): $\bar{Z}(x,\Phi) = \Sigma_{w \in V} Z(w,\Phi)\ \text{Prob}[v(x) = w]$, where the probability distribution over inheritance vectors here refers to the joint distribution over all pedigrees. To be precise, for a single pedigree we replace $S(v)$ by $\bar{S}$ in equation (3); the normalized scores for individual pedigrees are then combined into an overall score as in equation (4).

The only complication is in evaluating the statistical significance of $\bar{Z}$. Because $\bar{Z}$ is the expectation over the observed inheritance distribution, its statistical properties depend on the *distribution* of possible inheritance distributions (given the markers and pedigree structure). This distribution could be explicitly studied by Monte Carlo sampling from all possible realizations of the marker data. However, it is not hard to show that $\bar{Z}$ has the following properties under the null hypothesis of no linkage (see appendix D):

$$1. \qquad \text{mean}(\bar{Z}) = \text{mean}(Z) = 0\ ; \qquad (5)$$

$$2. \qquad \text{variance}(\bar{Z}) \leq \text{variance}(Z) = 1\ ; \qquad (6)$$

3. $\bar{Z}$ is asymptotically normally distributed as one studies a large number of similar pedigrees.

Moreover, $\bar{Z}$ approaches $\bar{Z}$ as information content approaches 100%, under both the null hypothesis of no linkage and the alternative hypothesis of linkage. Given these properties, it seems reasonable to evaluate the statistical significance of an observation $\bar{Z}_{obs}$ by using the null distribution of Z expected in the case of complete informativeness. The significance level is likely to be conservative (in view of 1 and 2; eqq. [5] and [6]) and becomes increasingly accurate as information content increases. We will refer to this approach as the *perfect-data approximation.*

Simulation studies (see below) show that the perfect-data approximation is indeed conservative but that it sacrifices relatively little power except when information content is very low. Indeed, significance levels appear to be within twofold of the empirical values obtained from simulations. The approximation should thus not hamper initial detection of interesting regions and should grow increasingly accurate as one genotypes additional markers in these regions.

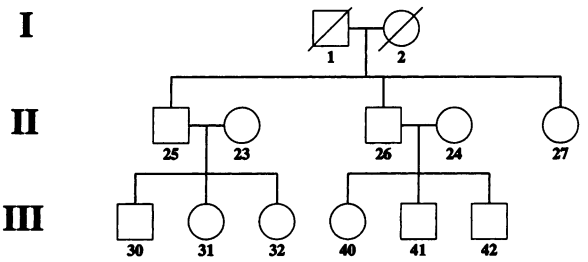


Figure 6 Pedigree structure used in the power simulations. Disease inheritance was simulated for four models: dominant, recessive, and two intermediate models (for details on the models, see table 1). Pedigrees were selected for analysis if they had at least three affecteds in generation III, including at least one affected individual in each sibship. Individuals in generation I were assumed to be unavailable for genotyping. Genotypic data were simulated for 11 markers spaced every 10 cM on a 100-cM map; all markers had five equally frequent alleles (heterogeneity .8). The disease locus was assumed to lie in the middle of the map, exactly at marker 6. A total of 100 pedigrees were simulated for each model, and 100 sets of 5 or 10 pedigrees each (depending on the model; see table 1) were resampled from this initial set for power calculations.

Implementation

We have implemented the calculation of both $\text{NPL}_{\text{pairs}}$ and NPL_{all} scores within GENEHUNTER. Given the inheritance distribution at each point in the genome, the calculation of NPL scores is rapid. The program reports the normalized score Z_i for each pedigree, the overall statistic Z, and the significance levels for the perfect-data approximation based on the exact approach.

Although we have focused only on S_{pairs} and S_{all}, it is straightforward to substitute other scoring functions to include further information about sharing among affected individuals or even about nonsharing between affected individuals and unaffected individuals. Such scoring functions can be easily incorporated into GENEHUNTER.

Evaluation of NPL Analysis

Power Comparisons

We compared the performance of the various linkage methods on simulated data, assuming dominant, recessive, and two intermediate models and a 10-cM genetic map with markers having heterozygosity of 80%. The pedigree structure used in the simulations is shown in figure 6 (for details, see the legend to fig. 6). The pedigrees were analyzed by using complete multipoint parametric linkage analysis (under the model used to generate the data), complete multipoint NPL analysis (using both the S_{pairs} and S_{all} scoring functions), and APM analysis (Weeks and Lange 1988, 1992). The performance

Table 1

Power Comparisons Based on Simulations

	Power to Detect Linkage with N Pedigrees (%)					Expected No. of Pedigrees Required for Detection of Linkage				
	p =					p =				
	.05	.01	.001	.0001	.00001	.05	.01	.001	.0001	.00001
Dominant:										
NPL_{all}	100	99	95	92	81	1	1	2	3	4
NPL_{pairs}	99	95	67	33	10	2	3	5	7	9
APM	62	42	15	5	2	5	8	15	22	29
LOD	100	100	100	100	100	1	1	2	2	3
Recessive:										
NPL_{all}	100	99	96	79	66	1	2	3	3	4
NPL_{pairs}	100	100	97	82	62	1	2	3	4	5
APM	87	59	34	19	10	2	4	6	9	11
LOD	100	100	100	100	100	1	1	2	3	3
Complex 1:										
NPL_{all}	64	40	17	7	4	10	20	34	50	65
NPL_{pairs}	58	25	7	1	0	16	31	55	79	102
APM	40	23	10	2	1	21	42	74	107	141
LOD	77	57	27	4	1	6	11	19	28	36
Complex 2:										
NPL_{all}	100	99	98	92	79	2	3	4	6	8
NPL_{pairs}	100	99	92	71	49	2	4	6	8	11
APM	83	68	41	14	5	4	8	14	20	26
LOD	99	99	97	95	87	1	2	4	5	7

NOTE.—N = 5 was used for the dominant and recessive models, and N = 10 was used for the two complex models. Power was defined as the number of data sets (from 100) in which the appropriate threshold was exceeded. Model parameters were as follows (f_d = disease gene frequency; p_{++}, p_{+d}, and p_{dd} = penetrances of ++, +d, and dd genotypes, respectively): for the dominant model, f_d = .01, p_{++} = .001, p_{+d} = .999, and p_{dd} = .999; for the recessive model, f_d = .05, p_{++} = .001, p_{+d} = .001, and p_{dd} = .999; for the complex model 1, f_d = .05, p_{++} = .05, p_{+d} = .4, and p_{dd} = .6; and, for the complex model 2, f_d = .01, p_{++} = .01, p_{+d} = .45, and p_{dd} = .75. These parameters correspond to disease incidence of 2.1%, 0.35%, 8.5%, and 1.9%, respectively, and to phenocopy rates of 5%, 29%, 53%, and 52%, respectively. The thresholds used for asymptotic significance levels of .05, .01, .001, and .0001, and .00001 were .59, 1.17, 2.07, 3.00, and 3.95, respectively, for the LOD score (LOD) and 1.65, 2.33, 3.09, 3.72, and 4.27, respectively, for the normal scores (NPL and APM). LOD scores were computed by using GENEHUNTER, in order to carry out multipoint analysis with 11 markers in reasonable time. Multipoint NPL statistics and multipoint LOD scores were computed by using all 11 markers simultaneously. As noted in the text, multilocus APM does not compute a *multipoint* statistic as a function of location. Instead, a statistic testing linkage to a region is computed. Recombination between loci is not fully taken into account, with the result that the statistic can decrease as additional flanking markers are considered, even in the presence of linkage. Therefore, we computed single-locus APM statistics by using the marker at the true locus, as well as multilocus APM statistics including 1, 2, 3, 4, and 5 closest flanking markers on each side, and we chose the highest statistic for each replicate when estimating power.

of the parametric LOD-score method under the correct model was used as a benchmark, although it should be noted that the correct model is usually unknown and that model misspecification can lead to considerable loss of power.

We used two criteria to assess performance: (1) the power to detect a locus in a fixed sample of pedigrees and (2) the expected number of pedigrees required to detect a locus. Both measures were completed for various nominal significance levels (p = .05, .01, .001, .0001, and .00001). The results are summarized in table 1. Three main conclusions emerge.

First, the NPL_{all} statistic performed better than the NPL_{pairs} statistic in all cases studied (except for the recessive case, where the two showed comparable performance). This accords with the intuition that testing whether the same allele is found IBD in many affected relatives is a more powerful strategy than considering one relative pair at a time. The NPL_{all} statistic thus appears to have the desirable property of robustness, and it was used in all other comparisons.

Second, the NPL statistic was much more powerful than the APM statistic, for all models examined. On average, at a given significance level, NPL required two to seven times fewer pedigrees for detecting linkage. The greater power of NPL is explained by its efficient use of all available information from simultaneous consideration of both all relatives and all markers.

Table 2

Empirical False-Positive Rates Observed in 50,000 Simulations

	False-Positive Rate at p =				
	.05	.01	.001	.0001	.00001
S_{all}, exact	.03	.004	.0003	.00002	0
S_{all}, normal	.04	.008	.001	.0002	.00006
S_{pairs}, exact	.03	.005	.0004	0	0
S_{pairs}, normal	.04	.008	.001	.0001	0

NOTE.—Genotypes for 50,000 data sets consisting of seven small two- and three-generation pedigrees were simulated, under the assumption that there is no linked disease-causing locus. "exact" refers to p values obtained from the exact distribution of scores; and "normal" refers to p values obtained from the normal approximation. The perfect-data approximation is used in both cases.

Third, the performance of NPL was roughly comparable to that of the LOD-score analysis under the correct model. NPL_{all} thus appears to provide a nonparametric pedigree-analysis method that loses relatively little power when compared with the best parametric method. This feature is particularly significant because the NPL method requires neither consideration of multiple models of inheritance (thus avoiding corrections for multiple testing) nor advance knowledge of the correct model of inheritance (thus avoiding problems of misspecification).

In addition to the power comparisons, we also examined the false-positive rate of NPL via simulation (table 2). The theoretical significance levels based on the perfect-data approximation provide a somewhat conservative test (with empirical false-positive rates roughly half those expected from theory), whereas those based on the asymptotic approximation of normality are closer to (and occasionally exceed) the empirical values. In summary, the procedures for evaluating statistical significance that are outlined above appear to be reasonable.

Application to Idiopathic Generalized Epilepsy

To compare NPL and APM on real data, we reanalyzed pedigrees with idiopathic generalized epilepsy (IGE) that recently were reported by Zara et al. (1995). IGE is a neurological disorder of unknown etiology characterized by recurring seizures. The pedigrees are shown in figure 7. Zara et al. used APM to obtain evidence for linkage of IGE to chromosome 8q24. Single-locus APM gave the strongest evidence for linkage at D8S256, with an APM statistic of 3.44, when allele frequencies taken from GDB were used, and 2.90, when allele frequencies estimated from the study sample were used. (We quote only the APM scores obtained with the $1/\sqrt{p}$ weighting function, which gave the strongest results). These APM scores correspond, respectively, to theoretical p values of .0003 and .002 when the statistic is assumed to follow a normal distribution and to empirical p values of .002 and .006 on the basis of simulations (Zara et al. 1995). Multilocus APM using D8S284, D8S256, and D8S534 gave somewhat weaker evidence for linkage. The statistics were 2.647 (GDB allele frequencies) and 1.478 (sample allele frequencies), corresponding to theoretical p values of .004 and .018, respectively, and to empirical p values of .008 and .07, respectively. Zara et al. considered these results as suggestive of the presence of an IGE susceptibility locus on 8q24 and stressed the need for confirmation in additional family sets.

We reanalyzed these data, using the NPL statistic S_{pairs}, which is the appropriate IBD generalization of the IBS APM statistic. A single-marker analysis yielded a score of 2.26 at D8S256 (p = .02). A complete multipoint analysis involving all three markers yielded a lower score, 1.79 (p = .063). The results were almost identical for both choices of allele frequencies.

Interestingly, NPL detects *less* evidence for linkage than does APM. Why? It turns out that the APM analysis gives weight to several instances of allele sharing that are IBS but not IBD. For example, it is clear that D8S256 is completely uninformative for linkage in family 13, since both parents are homozygous for the "10" allele (fig. 7). NPL assigns a score of 0 to this pedigree, since the allele sharing among affected individuals does not reflect IBD. In contrast, APM gives substantial weight to the observation of allele sharing at this locus. Indeed, the APM score for this pedigree is 1.08. Similarly, affected individuals 4 and 5 in family 8 share the "9" allele at D8S256 and thus contribute to the APM score. However, consideration of haplotypes clearly shows that this allele is not shared IBD. NPL analysis correctly does not detect this as sharing.

This example illustrates key advantages of NPL. Because NPL assesses IBD sharing on the basis of information from multiple markers and all genotyped relatives, it is less likely to be misled by chance sharing of alleles and is less sensitive to specification of allele frequencies. In contrast, APM has been reported to be prone to false positives, particularly when the single-locus method is used and correct allele frequencies are not known (Babron et al. 1993; Weeks and Harby 1995).

Application to Schizophrenia

To further evaluate the performance of NPL, we applied it to the data reported by Straub et al. (1995) on the 265 pedigrees in the Irish Study of High Density Schizophrenia Families (ISHDSF). This study used both parametric and nonparametric methods to map a potential schizophrenia-vulnerability locus to chromosome 6p24-22 and provided evidence for genetic heterogeneity. A total of 16 markers spanning 38.4 cM on 6p were examined.

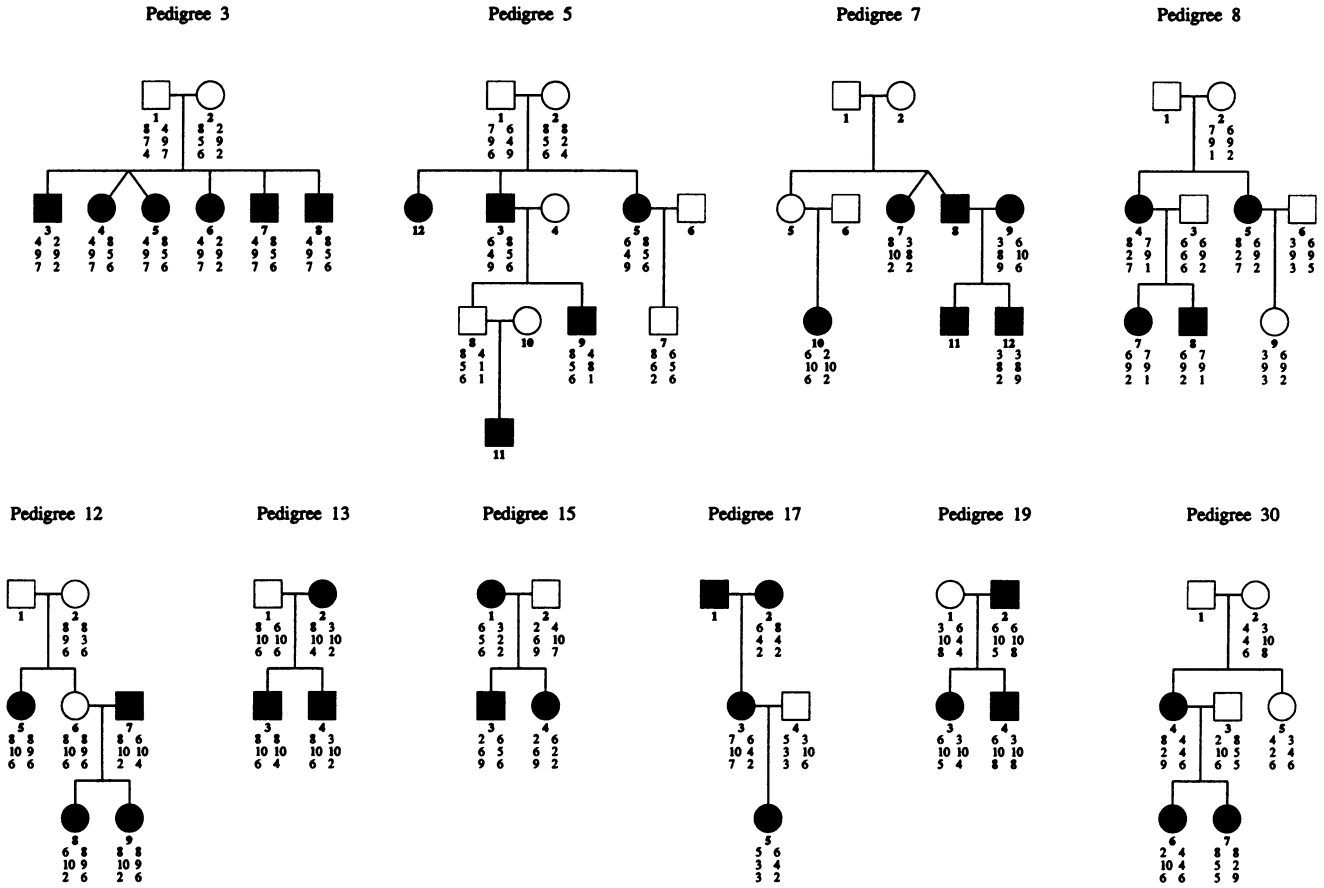


Figure 7 Ten pedigrees used in the IGE study, redrawn from the report by Zara et al. (1995). Blackened symbols indicate individuals considered as affected in the analysis. Genotypes for D8S284, D8S256, and D8S534 (from top to bottom) are shown under each individual's symbol.

In their parametric analysis, Straub et al. computed two-point LOD scores under the assumption of heterogeneity, using four genetic models each with four diagnostic categories. They obtained the strongest evidence for linkage at marker D6S296, under the "Pen" model and the broad diagnostic category (D1–D8), with a two-point LOD score of 3.51 (p = .0002) at θ = .004 and proportion of linked pedigrees α = .40. To extract all available linkage information, we extended the parametric study from single-marker analyses to a complete 16-marker multipoint analysis under the same model and diagnostic category (fig. 8*A*). The LOD curve peaked at D6S470 (2.6 cM proximal of D6S296), with a LOD score of 2.96 and α = .26. The multipoint results are likely to be more accurate because they do not rely on properties of a single marker. Indeed, the estimate of the heterogeneity parameter α agrees more closely with the estimate of 15%–30% that Straub et al. (1995) obtained by other means. Multipoint analysis also allowed us to compute the two-LOD support region, which extended over a 24-cM interval from D6S477 to D6S422.

Straub et al. also performed nonparametric single-marker analysis with ESPA (extended-sib-pair analysis [Sandkuijl 1989]), but they obtained much weaker evidence of linkage. Under the broad diagnostic category, they found values of p = .17 at D6S296 and p = .03 at D6S285, which is 16 cM proximal of D6S296. (Under the narrower D1–D5 diagnostic category, a p value of .005 was found at D6S285.) To compare the power of NPL analysis, we performed a complete 16-marker analysis with the NPL_{all} statistic. Under the broad diagnostic category, we obtained a maximum NPL score of 3.25 (p = .0005) at D6S470, in the same position as that of the multipoint LOD-score peak (fig. 8*B*). Secondary peaks of ~2.9 were obtained at D6S260 and D6S422.

Our results support the findings by Straub et al. (1995) and illustrate the advantages of the new analysis method. NPL analysis provided the same degree of evidence for linkage as did the parametric LOD-score method, without the need to examine multiple models of inheritance. (Of course, the choice of diagnostic categories remains important.) NPL provided strong evidence for linkage, whereas the other nonparametric method, ESPA, showed, at best, weak evidence. We at-

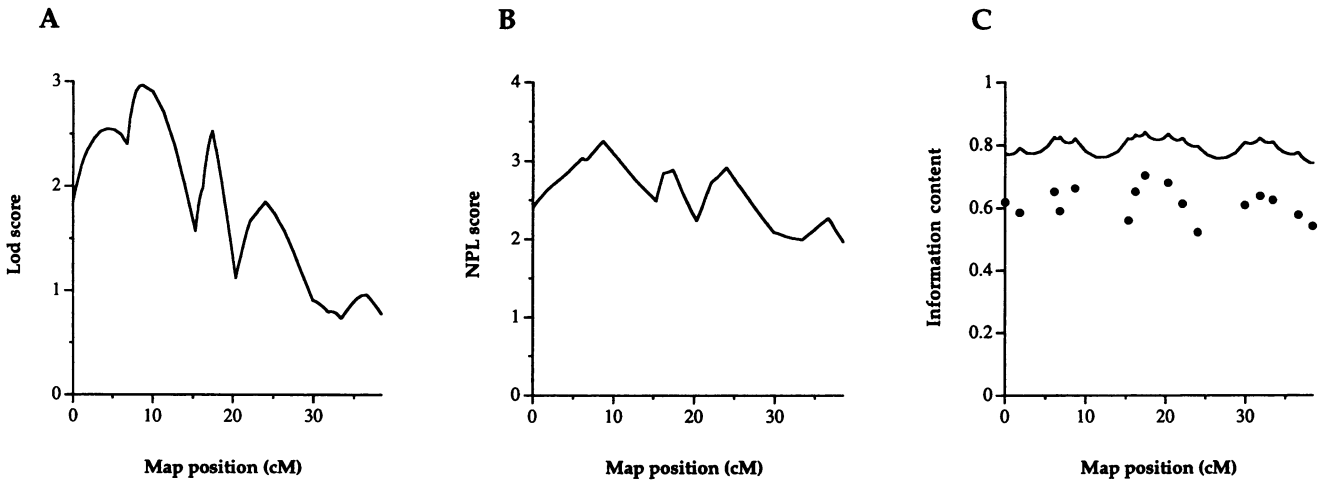


Figure 8 Analysis of pedigree data from the Irish schizophrenia study by Straub et al. (1995). Map position is in Kosambi centimorgams, from D6S477. *A*, Multipoint LOD scores computed under the Pen model, against a fixed map of 16 markers on chromosome 6p (LOD scores were computed at all markers and at four points within each interval between markers). *B*, Multipoint NPL scores (S_{all} statistic). *C*, Information-content mapping. The solid line shows the multipoint information content across the map when all 16 markers are used simultaneously; and black dots show the information content of individual markers.

tribute the superior performance of NPL to its efficient use of information from multiple markers and from all family members. To examine this point, we calculated the information content for single-marker analysis versus that for the complete multipoint analysis (fig. 8*C*). Whereas the information content of *individual* markers ranged from ~50% to 70%, the information content for the *map* of markers was ~80% across the entire region, indicating a substantial gain in informativeness for the multipoint approach. In summary, the various analyses indicate that NPL provides a useful, powerful, and robust method for demonstrating linkage to a disease with an uncertain mode of inheritance in a heterogeneous data set.

Haplotype Determination

Last, we turn to the problem of inference of haplotypes—that is, the determination of the particular founder alleles carried on each chromosome. It is often useful to infer haplotypes in order to identify double crossovers that may reveal erroneous data, to visualize single crossovers that may help confine a gene hunt, and to seek evidence of an ancestral chromosome in an isolated population. Some systematic methods for haplotype reconstruction that have been suggested previously have been based on rule-based heuristics, approximate-likelihood calculations, or exact-likelihood calculations (Weeks et al. 1995). The first two methods are ad hoc and sometimes fail to produce the best haplotype reconstruction, particularly in the presence of missing data. The third method has hitherto been limited to small numbers of markers and pedigrees with few missing data, owing to computational problems (Weeks et al. 1995).

Inheritance vectors provide a general framework for haplotype reconstruction. Since the inheritance vectors completely determine the haplotype, the problem reduces to choosing the "optimal" inheritance vector at the loci to be haplotyped. In the HMM literature, this is a well-studied question known as the "hidden-state reconstruction problem." There are two standard solutions, based on somewhat different optimality criteria (Rabiner 1989):

1. The first approach is to treat each locus separately and select the most likely inheritance vector at each locus (i.e., such that $P_{\text{complete}}(\boldsymbol{v})$ is largest). This method is clearly trivial to implement, given the inheritance distribution.
2. The second approach is to treat the loci together and select the most likely *set* of vectors at the loci (i.e., the vectors having the largest joint probability when considered as a sample path of the underlying Markov chain). This can be accomplished by using the Viterbi algorithm (Rabiner 1989).

The first method has the advantages that it is simple and easily reveals regions of uncertainty (in which distinct vectors have similar probabilities). The second method has theoretical appeal because it finds the globally most likely inheritance pattern. In practice, we find that both approaches tend to yield similar performance and results.

We have implemented both methods for haplotype reconstruction within GENEHUNTER. The program

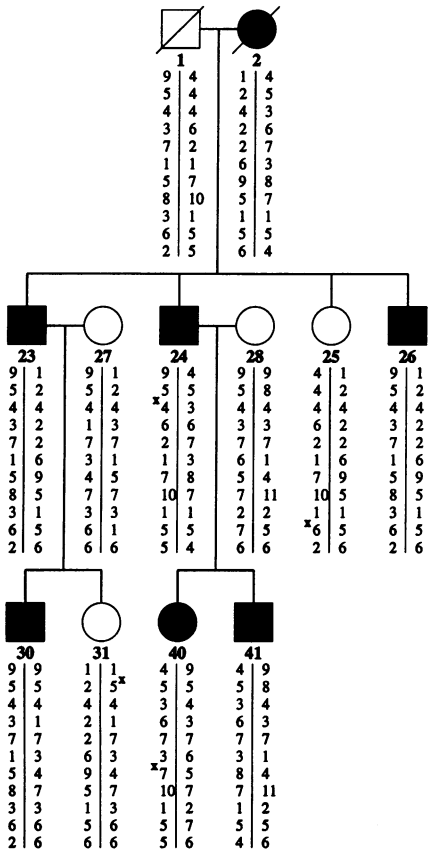


Figure 9 Example of haplotype reconstruction. Eleven markers were spaced every 2 cM across a 20-cM region. Individuals in generation 1 were not genotyped. Crossovers are denoted by x's.

reports the most likely haplotype, marking crossovers and highlighting double crossovers. It also indicates regions of haplotype uncertainty. The methods work for any number of markers and find the most likely haplotype even in pedigrees with missing data. An example is shown in figure 9.

Discussion

Collecting and genotyping data sets of a size sufficient for mapping complex disease genes is a formidable task, and it is desirable to have powerful analysis tools that efficiently make use of all available data. This requires (*a*) multipoint methods that extract inheritance information from all markers and (*b*) robust, nonparametric linkage methods that take account of all pedigree members. Currently available methods fall short in both regards.

In this paper, we describe algorithms for extracting all available inheritance information about segregation at every point in the genome, based on genotypes at any number of markers considered simultaneously. Such multipoint analysis is important for several reasons: (i) Even with relatively polymorphic microsatellite loci, multipoint analysis of many markers is required to infer IBD across several generations in pedigrees with substantial missing data; it can thus substantially increase the power to detect linkage. (ii) Multipoint analysis is more robust to misspecification of allele frequencies and statistical fluctuations at individual markers and can provide a confidence interval for the location of the gene. (iii) We expect that human genetic studies will soon employ a third-generation genetic map consisting of bi-allelic markers, because such markers are potentially amenable to high-throughput automation; their lower degree of informativeness can be offset through the use of a somewhat denser map, but this will depend crucially on the ability to perform extensive multipoint analysis.

The available inheritance information is captured in the multipoint inheritance distribution. This distribution provides a natural definition of information-content mapping, which measures the extent to which all inheritance information has been extracted. In addition, the inheritance distribution allows reliable reconstruction of many-marker haplotypes, even in pedigrees with missing data.

The inheritance distribution provides a unified framework for both parametric and nonparametric analysis. We have shown how to apply it to perform multipoint parametric LOD-score calculations and to define a multipoint nonparametric method, NPL. The framework also makes it straightforward to incorporate other linkage statistics.

We have studied the performance of NPL in applications to both simulated and actual data. NPL appears to have many advantages over the commonly used APM method, including much greater power to detect linkage and less sensitivity to misspecification of allele frequencies. In fact, in our comparisons, NPL was nearly as powerful as LOD-score analysis under the correct parametric model—but without the need to know and specify the model in advance. Because it appears to be robust to uncertainty about mode of inheritance and to lose little power compared with parametric methods, NPL would seem to be the method of choice for linkage analysis of pedigree data for complex traits. Of the two NPL methods described, we favor the NPL_{all} statistic and recommend using the perfect-data approximation to the significance level, for both the exact and the normal distributions. (The exact distribution provides a more accurate estimate of significance.)

The methods described in this paper are computationally feasible for pedigrees of moderate size ($2n - f \leq 16$, with current workstations), although not for large multigenerational pedigrees of the sort used to map simple dominant disorders such as Huntington disease. Fortunately, moderate-sized pedigrees are precisely the kind of pedigrees being used in most complex-disease studies.

Such pedigrees are easier to collect for diseases characterized by late onset, low penetrance, and diagnostic uncertainty. They are also more likely to reflect the genetic etiology of the disease in the general population and are less likely to show intrafamilial genetic heterogeneity.

The methods described here have all been incorporated into a new interactive computer package, GENEHUNTER. The computer 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/genehunter) or from our World Wide Web site (http://www-genome.wi.mit.edu/ftp/distribution/software/genehunter). We hope that GENEHUNTER will ease the task of efficiently analyzing pedigree data from genetic studies of complex traits.

Acknowledgments

L.K. would like to thank David Haussler for useful discussions and the Aspen Center for Physics for its hospitality during the 1995 workshop on biological sequences. We thank Richard Straub, Kenneth Kendler, and colleagues for sharing data from ISHDSF; Massimo Pandolfo and colleagues for sharing their IGE data; and Melanie Mahtani, Elizabeth Widen, Mark McCarthy, Leif Groop, and other members of the Botnia diabetes project for helpful discussions and for sharing unpublished data. This work was supported in part by National Center for Human Genome Research grants HG00017 (to L.K.) and HG00098 (to E.S.L.).

Appendix A

Computing the Single-Locus Inheritance Distribution at Codominant Loci

We describe an algorithm for computing P_{marker}, the inheritance distribution at a codominant marker locus, conditional only on the data for that locus. We begin by noting that it is sufficient to calculate $P(\Phi_{\text{marker}}|v)$, the probability of observing the marker data for each inheritance vector v. One can then apply Bayes's theorem, together with the fact that all inheritance vectors are equally likely a priori, to calculate the probability distribution over inheritance vectors.

Let $X = \{x_1, x_2, \ldots, x_{2f}\}$ be symbols corresponding to the $2f$ founder alleles at the marker locus, which are assumed to be distinct by descent. An inheritance vector v specifies the precise founder alleles inherited by each individual in the pedigree; let $x_{i1}(v)$ and $x_{i2}(v)$ denote the alleles carried by the ith individual. Let $A = \{a_1, \ldots, a_k\}$ denote the observable allelic states; note that distinct founder alleles may have the same state. Let a_{i1} and a_{i2} be the two observed alleles carried by the ith individual.

An *assignment* of the founder alleles is a mapping function, $f: X \rightarrow A$. For any inheritance vector v, an assignment f is said to be *v-compatible* with the observed marker data if $\{f[x_{i1}(v)], f[x_{i2}(v)]\} = \{a_{i1}, a_{i2}\}$ for all individuals who have been genotyped. (In other words, the assignment of founder alleles specified by f and the transmission specified by v are compatible with the observed genotype data.) Let $p(a_i)$ denote the population frequency of alleles having state a_i. The *probability* of the assignment f is $\prod_i p[f(x_i)]$, which is the chance that the founder alleles will happen to have the states specified in the assignment.

For a given inheritance vector v, the quantity $P(\Phi_{\text{marker}}|v)$ is equal to the sum of the probabilities of all v-compatible assignments. It thus suffices to find all v-compatible assignments. This can be done through a simple graph-theoretic process. Given v, define a graph $G(v)$ whose vertices are the founder alleles $\{x_1, x_2, \ldots, x_{2f}\}$ and whose edges are $e_i = \{x_{i1}(v), x_{i2}(v)\}$, where i runs over all genotyped individuals. (Pairs of vertices in $G(v)$ can be connected by multiple edges.) Label edge e_i with the corresponding genotype $\{a_{i1}, a_{i2}\}$. The v-compatible assignments are those such that the label on each edge is consistent with the assignment of the two vertices of that edge. Choose an arbitrary starting vertex y. If y has no edges, then the corresponding founder allele does not appear in any genotyped individual, and it may be assigned to any a_i. If y has edges, then its assignment necessarily must lie in the intersection of the labels on all edges from y; there are thus, at most, two choices for y. Given the assignment of y, the assignment of each neighboring vertex z is uniquely determined (since the pair of assignments of y and z must correspond to the label on any edge connecting them). Similarly, assignments are uniquely determined for neighbors of neighbors of y, and so on. In other words, the assignment of y automatically forces the assignment of all other vertices in the same connected component. (If this process leads to no assignment conflicts, it produces the unique v-compatible assignment for the component, given the assignment of y. If it produces a conflict, there is no v-compatible assignment, given the assignment of y.) Each connected component can be treated separately.

For any given inheritance vector v, the running time is easily seen to be $O(n)$ to find all v-compatible assignments of graph $G(v)$ and thus to compute $P(\Phi_{\text{marker}}|v)$. The overall running time is thus $O(n2^{2n-f})$ to compute $P(\Phi_{\text{marker}}|v)$ for all v and to apply Bayes's theorem to calculate P_{marker}.

Appendix B

The REDUCE Algorithm for Fast HMM Computation: Second Speedup

As in the earlier description of the algorithm (Kruglyak et al. 1995), we identify the set of all n-bit binary vectors

with $(Z_2)^n$, the additive vector space over the field with two elements (i.e., vector addition is component-wise modulo 2). Switching the phase of the ith founder corresponds to addition of a vector s_i in which the bits representing that founder's meioses equal 1 and all other bits are 0. Let S denote the subspace spanned by the vectors $s_1, \ldots, s_f$. Equivalence classes of vectors that differ only by founder phase are precisely the cosets of S; each coset contains 2^f vectors. Because a pedigree contains no information about founder phase, vectors that differ only by founder phase have equal probability; that is, probability is constant on equivalence classes (cosets). The rows of the matrices $\mathbf{W}^k$ are also constant on cosets: $W^k_{\alpha,j} = W^k_{\beta,j}$, where α and β are two vectors in the same coset. We therefore can interpret probability vectors and $\mathbf{W}^k$ matrices as indexed by cosets rather than by vectors and can perform the matrix-reduction algorithm as before, with cosets replacing vectors. This reduces the complexity of the problem by a factor of 2^f (i.e., from 2^{2n} to 2^{2n-f}), resulting in comparable savings in both time and memory.

Appendix C

Computing the Single-Locus Inheritance Distribution at Disease Loci

We describe an algorithm for computing P_{disease}, the inheritance distribution at a disease-causing locus, conditional only on the phenotype data Φ_{disease} for the disease. The disease will be assumed to have an arbitrary but specified mode of inheritance and two alleles, normal and disease, of specified frequency. As in appendix A, we note that it is sufficient to calculate $P(\Phi_{\text{disease}} | v)$, the probability of observing the phenotypic data for each inheritance vector v.

We need to compute $P(\Phi_{\text{disease}} | v) = \Sigma_G P(\{g_i\} | v) \times \prod_i P(\Phi_i | g_i)$, where $P(\{g_i\} | v)$ is the joint probability of the genotypes $\{g_i\}$ of all pedigree members, conditional on the inheritance vector v, and where $P(\Phi_i | g_i)$ is the probability that the ith pedigree member has phenotype Φ_i, conditional on having genotype g_i (this probability is determined by the penetrance function for pedigree member i). Conditioning on the inheritance vector means that $\{g_i\}$ is completely determined by the founder genotypes. The sum over $\{g_i\}$ can be computed in three ways:

1. *Direct summation over founder genotypes.* With f founders, this requires computing $O(4^f)$ terms, which grows exponentially with the number of founders.
2. *Peeling in pedigrees without loops.* In peeling (Elston and Stewart 1971; Lange and Elston 1975; Whittemore and Halpern 1994*b*), one identifies *peripheral* nuclear families that are connected to the rest of the pedigree by a single individual, designated the *pivot.* One then computes the probability of the phenotype data in the nuclear family, conditional on the genotype of the pivot, and replaces the penetrance function of the pivot with these probabilities. In traditional peeling, one sums over all possible allele transmissions by the parents in the nuclear families. Conditioning on an inheritance vector specifies all transmissions; thus, only one term needs to be computed for each pair of parental genotypes. Peeling runs in time to $O(N_F)$, where N_F is the number of nuclear families in the pedigree.
3. *Loop breaking in pedigrees with loops.* If loops remain after all peripheral families have been peeled off, one can either (*a*) sum over the genotypes of all remaining founders, as in procedure 1, or (*b*) break loops, by creating loop breakers or "twins" and conditioning on their genotypes (Lange and Elston 1975; Whittemore and Halpern 1994*b*), and then peel off additional founders, as in procedure 2. Each founder who must be summed over in direct enumeration and each "twin" contribute a factor of 4 to the number of terms that must be computed; that is, the algorithm runs in time $O(N_F + 4^{r+t})$, where N_F is the number of nuclear families that can be peeled off, r is the number of founders summed over at the end, and t is the number of loop breakers. We minimize the running time by choosing to break or not break loops so that $r + t$ is as small as possible.

In pedigrees without loops, peeling runs in constant time for each inheritance vector and is very rapid. Although computing time increases for more-complex pedigrees, we consider only pedigrees of moderate size, which cannot contain both many loops and many founders. Therefore, in practice, we can rapidly compute $P(\Phi_{\text{disease}} | v)$—and hence LOD scores—for any pedigree for which the REDUCE algorithm is computationally feasible.

Appendix D

Distribution of $\bar{Z}$

We now prove equations (5) and (6), concerning the distribution of $\bar{Z}$ under the null hypothesis of no linkage. Let the operator E_G denote expectation over all possible realizations of the observed marker genotype data, G. Note that, for every inheritance vector w, $E_G[P(w | G)] = P_{\text{uniform}}(w)$ under the null hypothesis.

To prove equation (5), we note that $\bar{Z} = \Sigma_{w \in V} P(w | G) Z(w)$ for observed data G. Applying the operator E_G, we have

$$\begin{aligned}\text{mean}\,(\bar{Z}) &= \mathrm{E}_G(\bar{Z}) \\ &= \mathrm{E}_G\left[\sum_{w\in V} P(w|G)\,Z(w)\right] \\ &= \sum_{w\in V} \mathrm{E}_G[P(w|G)Z(w)] \\ &= \sum_{w\in V} P_{\text{uniform}}(w)Z(w) \\ &= \text{mean}(Z)\ .\end{aligned}$$

To prove equation (6), we note that

$$\left[\sum_{w\in V} P(w|G)Z(w)\right]^2 \leqslant \sum_{w\in V} P(w|G)[Z(w)]^2\ .$$

The inequality follows from a straightforward application of Jensen's inequality (Royden 1968), since $f(x) = x^2$ is a convex function. When the operator E_G is applied to both sides, the inequality becomes $\text{var}(\bar{Z}) \leqslant \text{var}(Z)$.

References

Babron MC, Martinez M, Bonaiti-Pellie C, Clerget-Darpoux F (1993) Linkage detection by the affected-pedigree-member method: what is really tested? Genet Epidemiol 10:389–394

Cannings C, Thompson EA, Skolnick MH (1978) Probability functions on complex pedigrees. Adv Appl Prob 10:26–61

Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J (1986) Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–399

Cottingham RW Jr, Idury RM, Schäffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252–263

Curtis D, Sham PC (1994) Using risk calculation to implement an extended relative pair analysis. Ann Hum Genet 58:151–162

——— (1995) Model-free linkage analysis using likelihoods. Am J Hum Genet 57:703–716

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

Guo S-W (1995) Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am J Hum Genet 56:1468–1476

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) Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454

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

Lange K, Elston RC (1975) Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. Hum Hered 25:95–105

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

Lathrop GM, Lalouel JM, White RL (1986) Construction of human linkage maps: likelihood calculations for multipoint analysis. Genet Epidemiol 3:39–52

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

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nat Genet 11:402–408

Ott J (1991) Analysis of human genetic linkage, rev. ed. Johns Hopkins University Press, Baltimore and London

Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77: 257–286

Royden HL (1968) Real analysis, 2d ed. Macmillan Publishing, New York

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

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27:379–423

Sobel E, Lange K (1993) Metropolis sampling in pedigree analysis. Stat Methods Med Res 2:263–282

Straub RE, MacLean CJ, O'Neill FA, Burke J, Murphy B, Duke F, Shinkwin R, et al (1995) A potential vulnerability locus for schizophrenia on chromosome 6p24-22: evidence for genetic heterogeneity. Nat Genet 11:287–293

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

Thomas A, Skolnick MH, Lewis CM (1994) Genomic mismatch scanning in pedigrees. Int Math Assoc J Math Appl Med Biol 11:1–16

Thompson EA (1994) Monte Carlo likelihood in genetic mapping. Stat Sci 9:355–366

Thompson EA, Wijsman EM (1994) Multilocus homozygosity mapping and autozygosity patterns. Am J Hum Genet Suppl 55:A167

Weeks DE, Harby LD (1995) The affected-pedigree-member method: power to detect linkage. Hum Hered 45:13–24

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

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

Weeks DE, Sobel E, O'Connell JR, Lange K (1995) Computer programs for multilocus haplotyping of general pedigrees. Am J Hum Genet 56:1506–1507

Whittemore AS, Halpern J (1994*a*) A class of tests for linkage using affected pedigree members. Biometrics 50:118–127

——— (1994*b*) Probability of gene identity by descent: computation and applications. Biometrics 50:109–117

Zara F, Bianchi A, Avnanzini G, Di Donato S, Castellotti B, Patel PI, Pandolfo M (1995) Mapping of genes predisposing to idiopathic generalized epilepsy. Hum Mol Genet 4:1201–1207