Genetic Linkage Analysis: an Irregular Statistical Problem

David Siegmund

Abstract. Linkage analysis, which has the goal of locating genes associated with particular traits in plants or animals (especially inherited diseases in humans), leads to a class of "irregular" statistical problems. These problems are discussed with reference to an idealized model, which serves as a point of departure for more realistic versions of the problem. Some general results, adapted from recent research into "change-point" problems, are presented; and more specific problems arising out of the underlying genetics are discussed.

1991 Mathematics Subject Classification: 62M40, 92D10
Keywords and Phrases: gene mapping, linkage analysis, change point, irregular

1. Introduction. The goal of gene mapping, or linkage analysis, is to locate genes that affect particular traits, especially genes that affect human susceptibility to particular diseases and also genes that affect productivity of agriculturally important species. An artificially simplified, but illuminating genetic model leads to the following class of statistical problems. Observations are available on a doubly indexed set of random variables $Z(c, i\Delta)$, where $c = 1, \cdots, 23$ indexes the set of human chromosomes of genetic lengths ℓ_c and $i\Delta, 0 \le i\Delta \le \ell_c$ are the locations of markers spaced at intermarker distance Δ along each chromosome. For different values of c the random variables are independent. For each fixed c, $Z(c,t)$ is a stationary Gaussian process in t, which satisfies

$$\mathrm{Var}[Z(c,t)] = 1, \quad \mathrm{Cov}[Z(c,s), Z(c,t)] = R(t-s). \tag{1}$$

A case of particular interest is $R(t) = \exp(-\beta|t|)$. For most or perhaps all values of c

$$E[Z(c,t)] = 0 \text{ for all } t, \tag{2}$$

while for some c', $0 < \tau < \ell_{c'}$ and $\xi > 0$

$$E[Z(c',t)] = \xi R(t-\tau). \tag{3}$$

The values of c', τ, and ξ are all unknown. Thus the data consist of a large number of zero mean Gaussian processes observed at equally spaced "time" points. A small number of these processes are superimposed on a mean value function defining a "peak" of an unknown height ξ at an unknown location τ, and having a known shape R. The statistical problems are to decide which chromosomes, if any, harbor such a location τ and estimate the location by a confidence region. These problems are "irregular" for two reasons: (i) the parameter τ is not identifiable when the nuisance parameter $\xi = 0$; the log likelihood function, which is proportional to $Z(c, \tau)$, is not a smooth function of the parameter τ, even if we are able to make continuous observations in t.

The purpose of this paper is (a) to explain briefly the genetic background of the preceding problems as they relate to mapping human disease genes, (b) propose a framework for their solutions that is useful as a point of departure for discussing more realistic versions of the problems, and (c) describe some alternative models designed to capture the complicating features arising in practice. Special consideration is given to the issue of multiple comparisons that arises through examining the large number of variables $Z(c, i\Delta)$ in searching for the relatively few values of c', t where the expected value is substantially different from 0, and to estimation of τ by confidence regions. Some of these problems can be understood in terms of recent literature on "change-point" problems, to which they are closely related.

2. GENETIC BACKGROUND. Given two related individuals, at a given locus in the genome two alleles are said to be identical by descent if they are inherited from a common ancestor. For example, a pair of half siblings can inherit zero or one allele identical by descent from their common parent, and according to Mendel's laws each of these possibilities has probability 1/2. Genes on different chromosomes segregate independently, while genes on the same chromosome tend to be inherited from the same parental chromosome, and are said to be *linked*. More precisely, if two half siblings share an allele identical by descent at locus t, they will share an allele identical by descent at a locus on a different chromosome with probability 1/2 and at a locus s on the same chromosome with a probability $(1-\phi) \in (1/2, 1)$. This probability is a decreasing function of the distance between s and t.

A pair of siblings can inherit zero or one allele identical by descent from their mother and similarly from their father, hence 0, 1, or 2 overall. For some purposes a single sib pair can be regarded as two independent half sib pairs, but in general siblings require a more complicated analysis. For ease of exposition, we consider only the much simpler case of half siblings.

The basic logic of linkage analysis is that if two relatives, e.g., half siblings or siblings, share an inherited trait, e.g., a disease, that is relatively rare in the population, it is likely that they share an allele predisposing them to the trait that has been inherited identical by descent. Thus the probability of identity by descent for an affected relative pair at a marker locus close to a trait locus is greater than the value given by Mendel's laws (1/2 in the case of half siblings). Our problem is to scan the genome of a sample of affected relatives in search of regions where the identity by descent exceeds the expected proportion by more than can be explained as a chance fluctuation.

A mathematical model for a pair of half siblings is as follows. Let X_t be 1 or 0 according as the half siblings are or are not identical by descent at locus t (on a chromosome c, *which henceforth is suppressed in the notation*). Then for a random pair of half siblings,

$$P\{X_t = 1\} = P\{X_t = 0\} = 1/2 \tag{4}$$

for all t; and for loci s and t on the same chromosome

$$P\{X_s = 1|X_t = 1\} = P\{X_s = 0|X_t = 0\} = 1 - \phi. \tag{5}$$

Assume that τ denotes a genetic locus predisposing to inheritance of the trait (and that there is no other trait locus on the given chromosome). Then for two half siblings sharing a trait in common,

$$P\{X_\tau = 1\} = (1 + \alpha)/2 > 1/2, \tag{6}$$

while the conditional probability (5) continues to hold for loci s, t on the same side of τ. In particular by taking $t = \tau$ in (5) we obtain $P\{X_s = 1\} = [1+\alpha(1-2\phi)]/2$. The value of ϕ in terms of the parameters s and t depends on the model used for the genetic process of recombination. According to the commonly used model suggested by Haldane in 1919,

$$\phi = [1 - \exp(-\beta|t - s|)]/2, \tag{7}$$

and more generally

$$\phi \sim \beta|t - s|/2 \text{ as } |t - s| \to 0. \tag{8}$$

The value of β is determined by the relation of the relative pair. For half siblings it is 0.04 when the units of genetic distance along a chromosome are centimorgans (cM). (One cM is defined as the distance for which the expected number of crossovers per meiosis is 0.01. The average length of a human chromosome is roughly 140 cM. See Suzuki *et al.* for a more thorough discussion.)

Assuming now that one observes identity by descent data for N independent half sibling pairs at marker loci, denoted $i\Delta$, equally spaced at intermarker distance Δ throughout the genome, we form the statistics

$$Z_{i\Delta} = N^{-1/2}\Sigma_{j=1}^{N}[2X_{i\Delta}^{j} - 1], \tag{11}$$

where the summation is over all half sibling pairs. It is possible starting from (11) to address the basic questions of Section 1 (cf. Feingold, 1993, Tu and Siegmund, 1998). A somewhat simpler and more complete analysis is possible if we introduce an additional approximation. It follows from the central limit theorem that as $N \to \infty$ and $\alpha \to 0$ in such a way that $N^{1/2}\alpha \to \xi \geq 0$ the process $Z_{i\Delta}$ defined in (11) converges in distribution to a process, which by (4)-(7) has the properties described in (1) - (3). By an abuse of notation we continue to denote this new process by $Z_{i\Delta}$. Thus we return to the problems already formulated in Section 1.

3. GENOME WIDE FALSE POSITIVE ERROR RATE. If $i\Delta$ in (11) is equal to τ, it follows from (6) that (11) is the score statistic for testing whether $\alpha = 0$; it is also the likelihood ratio statistic in the approximating Gaussian model. Since usually τ is unknown, to test for linkage somewhere on the genome we use

$$\max_c \max_i Z_{i\Delta}. \tag{12}$$

To evaluate approximately the false positive error rate, i.e., the probability under the hypothesis of no linkage throughout the entire genome that (12) exceeds a threshold b, we assume that $b \to \infty$ and $\Delta \to 0$, in such a way that $b\Delta^{1/2}$ converges to a positive constant. Then for a genome wide search

$$P\Big\{\max_c \max_i Z_{i\Delta} > b\Big\} \approx 1 - \exp\{-C\Big[1 - \Phi(b)\Big] - \beta L b \varphi(b) \nu\Big(b\{2\beta\Delta\}^{1/2}\Big)\}. \tag{13}$$

Here Φ and φ are the standard normal distribution function and density function, respectively, C is the number of chromosomes and $L = \Sigma_c \ell_c$ is the total length of the genome in cM. The function ν, which arises in the fluctuation theory of random walks developed by Spitzer in the 1950's, is defined by

$$\nu(x) = 2x^{-2} \exp[-2\Sigma n^{-1} \Phi(-x n^{1/2}/2)]. \tag{14}$$

For small x it is easily evaluated via the relation $\nu(x) = \exp(-\rho x) + o(x^2)$, where $\rho = -\zeta(1/2)/(2\pi)^{1/2} \approx 0.583$, while the series in (14) converges very rapidly for large x. For a numerical example, for markers every $\Delta = 1$ cM and a human genome of 23 chromosomes of average length 140 cM the threshold $b = 3.91$ gives a false positive error rate equal to the conventional 0.05. The approximation (13) was given by Feingold, Brown and Siegmund (1993), as an application of the method of Woodroofe (1976).

4. POWER. To obtain an approximation to the power that we detect a disease locus on a correct chromosome (for simplicity we assume there is at most one on any given chromosome), we first suppose that the disease locus τ is itself a marker locus. We then have the approximation

$$P\Big\{\max_k Z_{k\Delta} \geq b\Big\} \approx 1 - \Phi(b - \xi) + \varphi(b - \xi)\Big[2\nu/\xi - \nu^2/(b + \xi)^2\Big], \tag{15}$$

where $\nu = \nu\Big(b\{2\beta\Delta\}^{1/2}\Big)$, as defined above. The first term in (15) is simply the probability that the process exceeds the threshold b at the disease locus. A disease locus between marker loci needs a similar but more complicated argument involving the (correlated) process $Z_{i\Delta}$ at the two flanking markers. The resulting approximation requires a one dimensional numerical integration for its numerical evaluation.

For the 1 cM intermarker distance and threshold $b = 3.91$ considered in the preceding section, and a disease locus midway between two markers a noncentrality parameter of $\xi = 5.03$ is needed to achieve power of 0.9 to detect the disease locus.

For a given value of the genetic parameter α, this can be converted to a sample size requirement by virtue of the relation $\xi = N^{1/2}\alpha$.

5. Confidence regions. A confidence region can be used to identify a chromosomal region in which to concentrate the search for the exact location of a disease gene. We discuss here two methods that are motivated by the recent literature on "change-point" problems, which have essentially the same structure. These methods are (i) support regions and (ii) Bayesian credible sets. (See Siegmund, 1989, for a review of the change-point literature). Note that as a consequence of the irregularity of this problem, the maximum likelihood estimator of τ is not normally distributed, so it is not correct to use the maximum likelihood estimator plus or minus two estimated standard errors as an approximate 95% confidence interval.

We assume that a disease gene has been correctly identified to lie on a particular chromosome, which contains no other disease gene. For simplicity we assume that the locus τ is exactly a marker locus. Since many investigators type additional markers in the proximity of an apparent disease gene, this latter assumption is often approximately true in practice.

The traditional genetic technique for estimating the location of a disease gene is a support region, which for our purposes can be defined as follows. Given $c > 0$, a support region contains all loci $j\Delta$ such that

$$Z_{j\Delta}^2 \geq \max_i Z_{i\Delta}^2 - c. \tag{16}$$

Within the framework of the approximate Gaussian model, this is equivalent to the standard statistical technique of inverting the likelihood ratio test that $j\Delta$ is the disease locus, to obtain a confidence region. If the problem were regular, which in this case would require that Z_t be twice continuously differentiable in t, the probability of (16) would be given approximately by a χ^2 distribution with one degree of freedom; but that approximation is not correct here. By methods similar to those used to obtain (13) one can approximate the probability of (16) and show that (16) yields an approximate confidence region for the disease locus (Feingold, Brown and Siegmund, 1993, Lander and Kruglyak, 1995, Dupuis and Siegmund, 1998).

Because of the local linear decay near τ displayed in (3), the inequality (16) will be satisfied at all loci within a distance from τ of roughly $c/2\beta\xi^2$. Since ξ is proportional to $N^{1/2}$, the expected size of the support region is proportional to N^{-1}. This stands in contrast to regular problems, where the likelihood function decays quadratically, and the size of a confidence region is proportional to $N^{-1/2}$. It may be shown by more detailed analysis that a value $c \approx 4.5$ corresponds roughly to a 90% confidence interval when $\Delta = 1$ and $\beta = 0.04$. Then for $\xi \approx 5$, the value indicated above that one needs to detect linkage with power about 0.9, the expected size of a support region is about 5 cM. Since this corresponds to about 5×10^6, base pairs, one still needs additional information, invariably of a qualitatively different kind, to locate the gene with precision at the base pair level.

In his study of the closely related change-point problem, Cobb (1978) observed that if ξ were known, the problem of estimation of τ would have essentially the same structure as estimation of a location parameter. Hence Fisher's (1934) suggestion for estimating a location parameter, to use the conditional distribution of the maximum likelihood estimator given the ancillary statistic, in our case the local rate of decay of the likelihood function, is very attractive. Moreover, this suggestion has minimal computational requirements, since it can be effected by a formal Bayesian credible region based on a uniform prior distribution for τ. To accommodate unknown ξ, one can introduce a prior distribution for ξ or use the profile likelihood function obtained by maximization with respect to ξ for each fixed τ.

Dupuis and Siegmund (1998) have compared these two methods and find that they are roughly comparable, although the former is more robust under a variety of conditions.

6. Multilocus models. There are many additional problems that require a more detailed understanding of the underlying genetics than we have presented so far. In this section we discuss traits involving more than one gene, while in the next we very briefly point out several additional problems.

While some inherited human diseases are governed by a single gene, most of the more common ones having a genetic component, e.g., diabetes, breast cancer, Alzheimer's disease, are known or thought to involve multiple genes. Conceptually the simplest of these are *heterogeneous* traits, where susceptibility increases by virtue of a mutant allele at any one of several loci. It is, of course, possible that the genome scan defined above would identify several disease loci, even though there is no particular effort to do so. Typically a much larger sample size would be required than for a single gene trait having a comparable degree of heritability, since the evidence for linkage is divided among the different disease loci.

Three methods have been suggested to deal with heterogeneous traits: (i) simultaneous search, (ii) conditional search and (iii) homogenization. In simultaneous search, suggested originally by Lander and Botstein (1986), one hypothesizes a specific number, say two, trait loci and searches over combinations of putatitve loci to identify both simultaneously. Because there is a much larger number of multiple comparisons, a suitable threshold under the conditions assumed above would increase from the neighborhood of 4 to about 5 (in searching for two loci). Conditional search, which is appropriate after some trait loci have already been identified, involves stratification of the sample according to the identity by descent status at the (estimated) location of the already discovered loci in order to increase precision in searching for additional trait loci. See Dupuis, Brown and Siegmund (1995) for a theoretical analysis of these two methods. An interesting application of conditional search is contained in Morahan *et al.* (1996), who identified a gene on chromosome two for insulin dependent diabetes by conditioning on the identity by descent status of their sample of sib pairs at the HLA locus on chromosome 6, which had been implicated in several earlier studies.

A third approach to alleviate the problem of heterogeneity is to develop a narrow definition of the disease, in order to make the disease more homogeneous. In some cases this definition can be achieved statistically. A notable success was the identification of the breast cancer gene BRCA1 by defining the trait to be early onset breast cancer. A recent attempt in the same direction involved a search for a gene contributing to noninsulin dependent diabetes (Mahtani *et al.*, 1996). After failing to find evidence of linkage in the complete study group, the pedigrees in the study were identified with their average level of a quantitative covariate thought to be associated with the trait. The analysis was repeated with only those pedigrees in the most extreme 25% of the distribution of this covariate, then the most extreme 50%, then the most extreme 75%. The genome scan in the most extreme 25% turned up a value that would have been marginally significant at the 0.05 level if the phenotype had been defined *a priori*, but now there is the second dimension of multiple comparisons (i.e., the search over levels of the covariate) to account for.

A suitable model to analyze this two dimensional search within the Gaussian framework introduced above is as follows. Let $Z(t,k)$ for $k = 1, \cdots, m$ be independent identically distributed Gaussian processes in t as defined in Section 1. Here k denotes levels of the covariate and for convenience is assumed to involve equal quantiles of its distribution. Then let

$$S(t,k) = k^{-1/2}\Sigma_{i=1}^{k} Z(t,i).$$

Linkage is detected if

$$\max_{1\le k\le m} \max_{c} \max_{j} S(j\Delta, k) \ge b \tag{17}$$

for a suitable threshold b. Using the method of Siegmund (1988), which generalizes Woodroofe (1976) to multidimensional time, one finds under the hypothesis of no linkage that the probability of (17) is approximately

$$1 - \exp\Big(-\beta L\nu[b(2\beta\Delta)^{1/2}]b^3\phi(b)\int_{bm^{-1/2}}^{\infty} x^{-1}\nu(x)\; dx\Big). \tag{18}$$

For the threshold $b = 3.91$ appropriate for the simple scan of Section 1 when $\Delta = 1$, we find when $m = 4$ that (18) is about 0.15. To obtain a false positive rate of 0.05, one must increase the threshold to $b = 4.2$. Some rough calculations, which should be more carefully analyzed, indicate that if the covariate is effective in "homogenizing" the original sample, one can sometimes achieve substantial gains in power after allowing for the increase in threshold.

In the paper of Mahtani *et al.* (1996) there was the additional problem that the study design required pedigrees to have at least three affecteds and employed a statistic whose distribution under the hypothesis of no linkage is skewed to the right. (See (iii) in Section 7 below.) As a consequence the p-value of their result was about 0.24 after one adjusts for skewness in addition to the two dimensional search.

7. Additional problems. Linkage analysis involves a large number of problems in addition to those discussed above. A few that have been the subject of recent research follow.

(i) The identity by descent data that form the basis of our previous discussion are intrinsically incomplete and require complicated algorithms to process. For example, for a given relative pair a particular marker may be "informative," so that we can observe the identity by descent status at that marker, or it may be "uninformative." Intermediate possibilities also exist. Since by (5) identity by descent status is correlated at nearby markers, it may be possible to infer that status at an uninformative marker from the status at nearby informative markers. For example, for half siblings it follows from (6) that the likelihood function (for the case of completely informative markers, when the trait locus τ is itself a marker locus) equals

$$\Pi_{j=1}^{N}\ (1+\alpha)^{X_\tau^j}(1-\alpha)^{1-X_\tau^j}.$$

Let G denote the observed genotypes of all individuals at all markers, and let P_0 denote probability under the hypothesis of no linkage. Then the likelihood function (relative to P_0) when some of the X_τ^j may not be observable is

$$\Pi_{j=1}^{N} E_0[(1+\alpha)^{X_\tau^j}(1-\alpha)^{1-X_\tau^j}|G] = \Pi_{j=1}^{N}[1+\alpha(2Y_\tau^j-1), \tag{19}$$

where $Y_\tau^j = E_0[X_\tau^j|G]$. Kruglyak *et al.* (1996) use hidden Markov chains to calculate the required conditional expectations. Their algorithm works best for a possibly large number of small pedigrees. Additional techniques are required for studies involving large pedigrees, which can make the required calculations extremely onerous (cf. Thompson, 1994). By differentiating (19) one sees that the score statistic for testing $\alpha = 0$ is

$$\hat{Z}_\tau = \Sigma_j[Y_\tau^j - 1/2]/[\Sigma_j \mathrm{Var}(Y_\tau^j)]^{1/2},$$

which reduces to (11) in the case of complete data. Since τ is unknown, we use $\max_c \max_i \hat{Z}_{i\Delta}$ to search the genome for evidence of linkage. By studying the correlation function of $\hat{Z}_{i\Delta}$, Teng and Siegmund (1998) show under certain conditions that a threshold b appropriate for the case of completely informative markers studied above is approximately correct for $\hat{Z}_{i\Delta}$ as well. They also study the effect of incompletely informative markers on the power to detect linkage. These problems are difficult, and pose a number of impediments to a completely satisfactory solution.

(ii) Many traits are defined by quantitative measurement rather than a yes/no dichotomy. Understanding the genetic basis of quantitative traits is also of interest in experimental genetics, e.g., for agriculturally important species or for animal models of human diseases. At the level of abstraction provided by Gaussian approximations one finds that linkage analysis of quantitative traits in humans and in experimental genetics has much in common with the problems discussed above, but many details are quite different–particularly when one considers various breeding designs available in experimental genetics (cf. Lander and Botstein, 1989; Dupuis and Siegmund, 1998).

(iii) The normal approximation suggested in Section 1 is adequate for the simple case of half siblings discussed there, because under the hypothesis of no linkage (11) is symmetrically distributed. In general, particularly when pedigrees contain more than two affecteds or distant affected relatives, the statistic is not symmetrically distributed and the normal approximation can be very poor. For example, for first cousins the probability of identity by descent at an arbitrary locus is $1/4$, so the statistic corresponding to (11) has a distribution skewed to the right; and the approximation (13) is anti-conservative. While it is possible to give approximations based directly on (11) or its analogue in more complex cases, these approximations can be onerous to evaluate numerically. A simple modification of (13) is given by Tu and Siegmund (1998). Let γ be the third moment of Z_t under the hypothesis of no linkage and $\theta = [-1 + (1 + 2b\gamma/N^{1/2})^{1/2}]/\gamma$. Then for a *single* chromosome of genetic length ℓ

$$P\{\max_{0 \le i\Delta < \ell} Z_{i\Delta} \ge b\}$$

$$\approx [1 - \Phi(b)] \exp(\gamma b^3/6N^{1/2}) + \nu\beta\ell b[2\pi(1+\gamma\theta)]^{-1/2} \exp[-N\theta^2(1+2\gamma\theta/3)/2], \quad (20)$$

where $\nu = \nu[b(2\beta\Delta)^{1/2}]$. Note that $\theta \sim b/N^{1/2}$ as either $N \to \infty$ or $\gamma \to 0$, and then (20) reduces to (13). An application of the analogous extension of (18) was described at the end of Section 6.

Acknowledgement. This research was partly supported by NSF Grant DMS-9704324 and by NIH Grant 5 R01 HG00898.

References

Cobb, G.W. (1978). The problem of the Nile: conditional solution to a change-point problem, *Biometrica* 62, 243-251.

Dupuis J., Brown P., Siegmund D. (1995). Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent. *Genetics* 140, 843-856

Dupuis J. and Siegmund, D. (1998). Statistical methods for mapping quantitative trait loci from a dense set of markers, submitted for publication.

Feingold E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J. Appl. Probab.* 30, 766-779

Feingold, E., Brown, P.O., Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent, *Am. J. Hum. Genetics*, 53, 234-251

Fisher, R. A. (1934). Two new properties of mathematical likelihood, *Proc. Roy. Soc. A* 144 285-307.

Griffiths, A.J.F., Miller, J.H., Suzuki, D.T., Lewontin, R.C., Gelbart, W.M. (1996). *An Introduction to Genetic Analysis*, 6th edition, W.H. Freeman and Company, New York.

Kruglyak L., Daly M.J., Reeve-Daly M.P., Lander E.S. (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *The American Journal of Human Genetics* 58, 1347-1363.

Kruglyak, L. and Lander, E.S. (1995). High-resolution genetic mapping of complex traits, *Am. J. Hum. Genet.* 56, 1212-1223.

Lander, E.S. and Botstein, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphis, *Proc. Nat. Acad. Sci. USA* 83, 7353-7357.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps, *Genetics* 121, 185-199.

Lander, E. S. and Schork, N.J. (1994). *Genetic Dissection of complex traits*, Science 265, 2037-2048.

Mahtani, M.M., Widen, E., Lehto, M., Thomas, J., McCarthy, M., Brayer, J., Bryant, B., Chan, G., Daly, M., Forsblom, C., Kanninen, T., Kirby, A., Kruglyak, L., Munnelly, K., Parkkonen, M., Reeve-Daly, M.P., Weaver, A., Brettin, T., Duyk, G., Lander, E.S. and Groop, L.C. (1996). Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families, *Nature Genetics* 14, 90-94.

Morahan, G., Huang, D., Tait, B.D., Colman, P.G., and Harrison, L.C. (1996). Markers on distal chromosome 2q linked to insulin-dependent diabetes mellitus, *Science* 272, 1811-1813.

Risch, N. (1990a,b,c). Linkage strategies for genetically complex traits I, II, III. The power of affected relative pairs, *Am. J. Hum. Genetics* 46, 222-228, 229-241, 242-253.

Siegmund, D. (1988). Approximate tail probabilities for the maxima of some random fields, *Ann. Probab.* 16, 487-501.

Siegmund, D. (1989). Confidence sets in change-point problems, *International Statistical Review* 56, 31-48.

Teng, J. and Siegmund, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers, to appear in *Biometrics.*

Thompson, E. (1994). Monte Carlo likelihood in genetic mapping, *Statist. Sci.* 9, 355-366.

Tu, I.-P. and Siegmund, D. (1998). The maximum of a function of a Markov chain and application to linkage analysis, submitted for publication.

Woodroofe, M. (1976). Frequentist properties of Bayesian sequential tests, *Biometrika* 63, 101–110.

David Siegmund
Department of Statistics
Sequoia Hall
390 Serra Mall
Stanford University
Stanford, CA 94305
USA