# Gaussian Models for Genetic Linkage Analysis Using Complete High-Resolution Maps of Identity by Descent

E. Feingold,* Patrick O. Brown,†'‡'§ and D. Siegmund*

Departments of *Statistics, †Pediatrics, and ‡Biochemistry and §Howard Hughes Medical Institute, Stanford University

## Summary

**Gaussian-process models are developed to detect genetic linkage using complete high-resolution maps of identity by descent between affected relative pairs. Approximations are given for the significance level and power of the likelihood-ratio test of no linkage and for likelihood-ratio confidence regions for trait loci. The sample sizes required to detect linkage by using different classes of affected relative pairs are compared, and the problem of combining data from different classes of relatives is discussed.**

## Introduction

Classical linkage analysis in human genetics proceeds by evaluating the likelihood of the observed recombination pattern between a phenotype and a putative marker or a small number of markers, which, to be most useful, must occupy known map positions. The appropriate statistical theory is described by Ott (1991). In order to take advantage of increasingly precise RFLP linkage maps, Lander and Botstein (1986*a*, 1986*b*) have suggested methods for simultaneously testing linkage of Mendelian traits to an array of markers and, in a subsequent paper (Lander and Botstein 1989), discuss a method for searching the entire genetic map of an experimental organism, for quantitative-trait loci. A primary difficulty associated with this approach to linkage in humans is the need to collect large pedigrees spanning several generations, including family members with and without the trait, in order to obtain "informative" family units. Incomplete or age-dependent penetrance causes additional problems.

An alternative approach is to establish linkage through mapping the regions of genetic identity by descent of pairs of affected relatives, i.e., relatives who share a trait of interest. A notable advantage of this method is that only individuals having the trait need to be studied, and pedigrees can consist of as few as two affected relatives. In contrast to the well-documented large pedigrees required by classical linkage analysis, the method of affected relatives capitalizes on the possibility of using a large number of small pedigrees consisting only of affected individuals, who, at least for medically significant traits, are likely to be easily located and eager to cooperate. An illuminating discussion of the method of identity by descent for affected relative pairs is contained in the work of Risch (1990*a*, 1990*b*, 1990*c*). Risch, however, does not directly consider advantages that accrue if one can utilize a complete genetic map instead of isolated markers.

With a sufficiently large number of affected relative pairs, an analysis that maps regions of identity by descent along the entire genome can reveal the positions of genes that contribute even a slight susceptibility to the trait. The number required is a function of many variables, including the genetic relationships of the affected pairs, the relative risk (compared with the source population) of the trait in the individuals with a "susceptible" allele at the locus of interest, and the fraction of "affected" cases associated with a particular locus (which depends on both genetic heterogeneity and frequency of phenocopies or misdiagnoses). A critical determinant of the number of pairs required for this mapping strategy is the density and degree of polymorphism of the markers to which linkage is sought (Bishop and Williamson 1990; Risch 1990*c*). For practical purposes,

Received December 17, 1992; revision received February 24, 1993.

Address for correspondence and reprints: Dr. D. Siegmund, Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305.

0002-9297/93/5301-0027$02.00

the affected-relative-pair strategy requires analysis of a large number of highly informative polymorphic markers, closely spaced throughout the genome. Until recently, the lack of a practical laboratory method to satisfy this requirement has posed a barrier to widespread application of affected-relative-pair mapping of complex human traits.

Methods for typing a dense set of polymorphic markers are now approaching feasibility. Our approach is motivated by the particular laboratory method of genomic mismatch scanning (Nelson et al., in press), which can provide an essentially continuous specification of regions of identity by descent. The same statistical ideas are relevant to identity-by-descent data obtained from any polymorphic, reasonably dense mapped set of markers, e.g., an RFLP linkage map (Botstein et al. 1980).

In the present paper we discuss via a number of simple examples some Gaussian models for analyzing completely mapped identity-by-descent data. Questions to be addressed include the following: (1) How should we test for regions of enriched identity by descent, including the determination of thresholds that control the rate of false positives? (2) How can we determine the sample size required to give us a prescribed power to detect effects of a hypothesized magnitude? (3) When we believe that a region of enriched identity by descent has been detected, how do we specify the region over which subsequent, more careful searching should take place? (4) How do answers to the first three questions vary as a function of the density of the genetic map?

The paper is organized as follows: In the rest of this introduction we describe in the simplest possible context, where our data are derived entirely from a number of independent grandparent-grandchild pairs, a Markov-chain model that has been discussed in more detail by Feingold (in press). Next we develop a Gaussian approximation to the Markov-chain model and give some results for other relative pairs and combinations of classes of relatives. We also discuss briefly the complications that arise if regions of identity by descent are determined by reference to a limited fixed set of markers. Overall we find that the Gaussian models are substantially simpler to analyze than are the Markov-chain models. As a result, we can obtain more complete answers, which provide useful new insights. The unifying feature of our treatment is the systematic application of ideas from the recent statistical literature on change-point problems (e.g., see James et al. 1987; Siegmund 1989).

### 1. *The Markov-Chain Model*

Our basic experimental assumption is that, for any two relatives sharing a trait of interest, it can be determined where along their genomes the DNA sequences are identical by descent. For a given pair of relatives these regions of identity by descent can be described by a 0-1 process, say $Y_t$, where $Y_t = 1$ indicates identity by descent at a locus $t$ and $Y_t = 0$ indicates an absence of identity by descent at $t$. A gene contributing to a phenotype that is not common in the population but is shared by the relative pair is expected to be found in the region of identity by descent.

The data consist of a number of such 0-1 processes for various relative pairs sharing the phenotype of interest. The goal is to look for regions of identity by descent that are common to a large proportion of similarly affected relative pairs, with the purpose of discovering the location of the gene or genes contributing to the shared trait.

For simplicity we use the Haldane mapping function and assume a common rate of crossovers for male and female meioses. At the cost of some complication, these assumptions may be weakened.

*Remark.* To map rare recessive traits, it may be useful to analyze regions of homozygosity by descent of the offspring of consanguineous matings (Lander and Botstein 1987). The methods developed in the present paper can, with minor modifications, be applied to the statistical analysis of data from complete maps of homozygosity by descent. For a parent-child mating the associated Markov chain is slightly different from any that we encounter in identity-by-descent analyses. For most other matings the same process arises in analysis of identity by descent. For example, for the offspring of a mating of siblings the process describing the regions of homozygosity by descent is exactly the same as that encountered in identity-by-descent studies of first cousins, since the offspring of a mating of siblings is its own cousin. However, for studying homozygosity by descent, it seems less likely that large numbers of cases will be involved—and hence less plausible that the large-sample Gaussian models developed here will be directly applicable. Appropriate Markov-chain models will be discussed elsewhere (Feingold 1993).

### 2. *Grandparent-Grandchild Pairs*

Initially we consider the simplest possible case, where our data consist of a number of independent grandparent-grandchild pairs sharing the trait of interest. The meiosis that determines regions of identity by descent for any given pair is that of the intervening parent. Under the Haldane mapping function, which specifies that the number of crossovers as we move along each chromosome is described by a Poisson process, the probability of a recombination between two loci having genetic distances (in cM) $t$ and $s$ from the left end of the same chromosome is $\theta = [1 - \exp(-2\lambda|t - s|)]/2$, where $\lambda = 0.01$. Equivalently, the 0-1 process that indicates regions of identity by descent switches between the two states after independent exponentially distributed intervals having mean genetic length $1/\lambda$ cM. On any chromosome not containing a gene contributing to the trait of interest, at any locus $t$, the 0-1 process is found in each of its two possible states with probability 1/2 each.

Consider a chromosome containing a single locus $r$, at which one or more alleles confers susceptibility to the trait. In a given population suppose that grandparent-grandchild pairs having the trait have an increased probability $(1 + \alpha)/2$ of identity by descent at locus $r$. The parameter $\alpha$ measures the excess likelihood of identity by descent at $r$, among relative pairs selected for sharing the trait, but otherwise the definition is completely formal. For some purposes it is useful to have a genetic-epidemiological interpretation of $\alpha$. One possibility is to interpret $\alpha$ as the percentage of pairs having the trait on the basis of a shared allele at the locus $r$. We assume that the allele occurs sufficiently infrequently in the population that, with large probability, its occurrence in both members of a relative pair is due to identity by descent. The remaining $(1 - \alpha)100\%$ of pairs are assumed to have the trait for some reason unrelated to the locus $r$—and hence are assumed to share alleles identical by descent at $r$ with probability 1/2. A more sophisticated interpretation, arising from a detailed genetic model for the trait, is Risch's (1990*a*, 1990*b*) interpretation of $\alpha$ in terms of the increased risk of the trait appearing in a relative of a person who has the trait of interest. The parameter depends on both the trait and the relationship of the affected pair. For grandparent-grandchild pairs, in Risch's (1990*b*) simplest model $\alpha = (\lambda_O - 1)/(\lambda_O + 1)$, where $\lambda_O$ is the relative risk of an offspring of an affected parent to be affected, compared with the population prevalence of the trait. For a brief discussion, see Appendix B.

We now consider the process $X_t$, the number of grandparent-grandchild pairs, of a total of $N$ pairs, having identity by descent at the locus $t$. On each chromosome, $X_t$ is a Markov chain on the states $0, 1, \ldots, N$ whose transition rates are easily calculated from the assumed Haldane mapping function and the independence of the different relative pairs. On a chromosome not containing a locus contributing to the trait, the chain has a stationary distribution, which is binomial $(N, 1/2)$, so $X_t$ is usually found close to $N/2$. On a chromosome containing exactly one trait locus $r$, the expected value of $X_t$ at $t = r$ is $N(1 + \alpha)/2$, and, for general $t$, $N[1 + \alpha \exp(-2\lambda|t - r|)]/2$. Hence at loci $t$ close to $r$ we expect to find values $X_t$ rather greater than $N/2$, indicating a region of enriched identity by descent. It seems reasonable to look for trait loci in regions where $X_t$ takes on large values and to use the maximum value of $X_t$ as $t$ varies over each chromosome, as a statistic to test for the existence of such loci. If we use a threshold $a$ to determine the existence of such regions, the false-positive rate for each individual chromosome is

$$P\{\max_{0 \leqslant t \leqslant l} X_t \geqslant a\} , \tag{1}$$

where $l$ is the genetic length of the chromosome and the probability is computed under the assumption that there is no trait locus $r$. To apply this test simultaneously along an entire genome, we can use the independent assortment of chromosomes at meiosis to give an overall false-positive rate, which, for practical purposes, can be taken to be the sum of the false-positive rates for the individual chromosomes.

For a discussion of probability (1) and similar probabilities for data obtained from other relative pairs, see Feingold (in press). In general the processes involve functions of underlying, unseen Markov chains but are not themselves Markovian.

## Gaussian Approximations: Grandparent-Grandchild Pairs

We now turn to approximate Gaussian models. Initially we consider only the case of a large number of grandparent-grandchild pairs, and later we indicate by a number of examples the nature of the corresponding analysis for other affected relative pairs. It turns out that the Gaussian model described here is mathematically related to the model of Lander and Botstein (1989), which we therefore also discuss briefly.

### 1. *The Gaussian Model*

Suppose that, for the grandparent-grandchild model discussed above, the number of affected pairs, $N$, is large. The parameter $\alpha$ denotes the excess likelihood of identity by descent at the trait locus $r$. It is convenient to let $p$ denote the probability of identity by descent at an arbitrary locus, $p = 1/2$ for grandparent-grandchild pairs, and to introduce a new parameter defined by

$$\xi = N^{1/2}\alpha p \,. \tag{2}$$

Then, by the central limit theorem, for large values of $N$ the normalized statistic

$$(X_t - Np)/N^{1/2} \tag{3}$$

is approximately a Gauss-Markov process $Z_t$, which can be described as follows: Along each chromosome not containing the locus $r$, $Z_t$ is a stationary Ornstein-Uhlenbeck process with mean value 0 and covariance function $\sigma^2\exp(-\beta|t|)$, where $\sigma^2 = p(1-p) = 1/4$ and $\beta = 2\lambda$, with $\lambda$ being the crossover rate per unit of genetic distance $t$. When $t$ is in centimorgans, $\lambda = 0.01$. On the chromosome on which the distinguished locus $r$ resides, the process $Z_t$ is the same stationary Ornstein-Uhlenbeck process superimposed on the mean value function

$$\xi \exp(-\beta|t-r|)\,, \tag{4}$$

which has its maximum value equal to $\xi$ at the point $r$ and drops off exponentially as we move away from $r$. For basic definitions and for partial derivations of the results stated below, see Appendix A.

*Remarks.* (i) Although we restrict our explicit discussion to the case where alleles at a single locus may confer increased susceptibility to a trait, our long-range goals are to deal with polygenic traits. In that case we want to detect and estimate the location of all loci making significant contributions. When alleles at some or all of several loci may confer susceptibility to a trait, the enriched identity by descent at any one of these loci, as measured by the parameter $\alpha$, is likely to be small and hence difficult to detect without a large sample size. For example, in the Risch (1990*a,* 1990*b*) model for monogenic inheritance a value of $\lambda_O = 9$ corresponds to $\alpha = 0.8$. However, if two unlinked loci confer increased susceptibility via the multiplicative model of Risch (1990*a*), then $\lambda_O = \lambda_{1O}\lambda_{2O}$, where $\lambda_{iO}(i = 1{,}2)$ is a relative-risk factor associated with locus $r_i$. A similar Gaussian model applies, and the value $\alpha_i$ associated with $r_i$ is $(\lambda_{iO} - 1)/(\lambda_{iO} + 1)$. If the two loci contribute equally, so that $\lambda_{1O} = \lambda_{2O}$, then the same value of $\lambda_O = 9$ corresponds to $\alpha_i = 0.5$. A similar analysis of Risch's additive model yields $\alpha_i = 0.4$. In the numerical examples that follow we concentrate on relatively small values of $\alpha$. (ii) The specific form of the Haldane mapping function leads to the exponential mean value and covariance function for $Z_t$. If the fraction of recombinants between loci at genetic distance $t$ is given by $\theta = [1 - R(t)]/2$, then the mean value of $Z_t$ becomes $\xi R(t - r)$, while the covariance function is $R(t)/4$. The resulting theory is only slightly more complicated and is discussed below (see Proposition 4 of Appendix A). The main consequence of using a different mapping function would be a genome of a different total length, which one can see, from the results in the rest of the paper, would have only minor implications. (iii) For the model of Lander and Botstein (1989) dealing with quantitative traits in experimental genetics, this same Gaussian process describes the asymptotic behavior of the statistic (their notation) $n^{1/2}\beta^*(d)$. For a description of the behavior of the process when there is no quantitative-trait locus on the given chromosome, see their footnote A3. This corresponds to the special case $\xi = 0$ in our notation, but it is easy to show that the two models also coincide in the case when $\xi$ ($n^{1/2}b$ in their notation) is unequal to 0.

## 2. *Significance Level, Power, and Sample Size*

In what follows we discuss the process $Z_t$ for an arbitrary chromosome of genetic length $l$ and, to extend results to an entire genome, appeal to the independent assortment of chromosomes at meiosis. To study the process $Z_t$ on a given chromosome, we note that, although we observe the process over an interval $[0,l]$, where $l$ is the (genetic) length of the chromosome, the process can be assumed to be defined on the entire real line. As a consequence of Proposition 1 in Appendix A, we see that the log-likelihood function of the observed process $\{Z_t, 0 \leqslant t \leqslant l\}$ as a function of the unknown parameters $r$, $\xi$ equals

$$\sigma^{-2}[\xi Z_r - \xi^2/2]\,. \tag{5}$$

In particular, if $r$ were known, then a sufficient statistic for the remaining parameter $\xi$ would be $Z_r$. From formula (5) it follows that for testing the null hypothesis of no region of enriched identity by descent, $\xi = 0$, against the one-sided alternative, $\xi > 0$, the likelihood-ratio test rejects the null hypothesis if $\max_{0\leqslant t\leqslant l} Z_t/\sigma$ exceeds some threshold $b$. (This is equivalent to the test proposed above for the Markov-chain model.)

In the statistical literature there are several simple approximations for the significance level, $P_0\{\max_{0\leqslant t\leqslant l} Z_t/\sigma > b\}$, where the subscript 0 denotes that the probability is evaluated under the assumption that $\xi = 0$. For example, see Leadbetter et al. (1983, chap. 12), Siegmund (1985, chap. 4), and Aldous (1989, chap. D). For our calculations we have used the approximation

$$P_0\{\max Z_t/\sigma > b\} \approx 1 - \Phi(b) + \beta lb\phi(b)\,, \quad (6)$$

where $l$ is the length (in cM) of the chromosome and $\phi$ and $\Phi$ are the standard normal density and distribution functions, respectively. Although this approximation is not the best one for the present problem, it has the advantage that, with minor modifications, it is appropriate for problems in which the set of markers is discrete and for non-Markovian processes (see below). By the independent assortment of chromosomes, an overall significance level when this test is applied to each chromosome is approximately approximation (6) summed over all chromosomes.

It is also possible to give an approximation for the power of the test, i.e., the probability $P_\xi\{\max_t Z_t/\sigma > b\}$ for values of $\xi$ not equal to 0 and hence, via equation (2), to determine the number of affected pairs necessary to detect, with reasonably large probability, say .50 or .90, a deviation of given magnitude from the null hypothesis. By arguing along the lines of James et al. (1987), we obtain in Proposition 2 of Appendix A the following approximation for the power of the test:

$$P_\xi\{\max Z_t/\sigma > b\} \approx 1 - \Phi(b - \xi/\sigma) + \phi(b - \xi/\sigma)[2(\xi/\sigma)^{-1} - (\xi/\sigma + b)^{-1}]\,. \quad (7)$$

The first two terms on the right-hand side of approximation (7) give the probability that the process at the trait locus, $Z_r/\sigma$, exceeds $b$. The final expression is an approximation for the probability that the process is below the threshold at the trait locus $r$ but, because of random variation, exceeds the threshold at some nearby locus. The approximation (7) is predicated on the assumption that the locus $r$ is not too close to either end of a chromosome. When $r$ is at an end of a chromosome, there is slightly less power, and the term in square brackets in approximation (7) is just $(\xi/\sigma)^{-1}$. Note that approximation (7) depends critically on the *noncentrality* parameter $\xi/\sigma$. It depends indirectly on the value of $\beta$, to the extent that the value of $b$ needed to achieve a desired false-positive rate depends on $\beta$ (see approximation [6]).

For a numerical example, we consider the mythical unicorn, which we will suppose to have 25 pairs of chromosomes, each of which is 100 cM in length. The numerical results would be almost the same for humans having 23 pairs of chromosomes of a total genetic length in the range 3,000–3,600 cM. To obtain a significance level of about .002 for each chromosome—and hence an overall significance level of about .05—we can, according to approximation (6), take $b = 3.84$. This corresponds to a threshold of 3.2 on the LOD scale, which is traditionally used with a threshold of 3 in classical linkage analysis (Ott 1991). In terms of the process $X_t$, which counts the number of cases (of the total of $N$ affected pairs) of identity by descent at the locus $t$, by statistic (3) and approximation (6) the threshold is $Np+\sigma bN^{1/2} = 50 + (0.05)(3.84)(10) = 69.2$ for $N = 100$, in agreement with Feingold (in press). According to approximation (7), 50% power is achieved at $\xi = 1.72$. Using equation (2), we see that we can achieve 50% power to detect a gene responsible for 50% of the cases of a given trait, with a sample size of $N = 48$ affected pairs. If we want 90% power to detect a single locus contributing to 50% of the cases of a trait, we need about $N = 90$ affected pairs. This later result appears to be roughly consistent with a calculation by Risch (1990*b*), if we equate our parameter $\alpha$ to his $(\lambda_O - 1)/(\lambda_O + 1)$. However, Risch's calculation is concerned with a single marker assumed to lie at zero recombination distance from the trait locus of interest. This will rarely be the case when linkage to individual markers is being tested.

*Remarks.* The following technical mathematical features of the Gaussian approximation are worth noting. For $r \leqslant s \leqslant t$ the exact covariance of the process (3) is $\sigma^2[\exp(-\beta|t - s|) - \alpha^2\exp(-\beta\{t + s - 2r\})]$. As $N \rightarrow \infty$, we assume that $\alpha \rightarrow 0$ in such a way that the parameter $\xi$ in equation (2) remains finite. As a consequence, the second term in the exact covariance function of process (3) converges to zero. The discrepancy between the exact and the asymptotic covariance functions will have some effect on the quality of the approximation (7), for the power of the test to detect linkage, especially if $\alpha$ is large. If we carry out a similar analysis with the more complicated, exact covariance function, we find that about $N = 83$ affected pairs are required for 90% power, instead of the $N = 90$ obtained above. Similarly, while it turns out that approximation (6) provides an excellent approximation in a number of cases, in general it can be improved by making a preliminary transformation of the process $X_t$. Since our main goal in this paper is to use the simplest possible models to obtain new insights, we defer to the future both a systematic quantitative assessment of the accuracy of our approximations and the possibility of improving them at the cost of some additional computation.

### 3. *Confidence Regions*

If, because of large values of $Z_t$ for a certain range of $t$, our test indicates the presence of a region or regions of enriched identity by descent, we would like to estimate the extent of the region likely to contain the unknown true values of $r$. Confidence regions provide an ideal statistical tool for this purpose. For a general discussion of this concept and its application to linkage analysis, see the work of Ott (1991, esp. secs. 3.6 and 4.4). The locus $r$ where the mean value of $Z_t$ reaches its maximum is a change point (for a definition and examples, see Appendix A), and consequently, to find a confidence region for $r$, one must go outside standard statistical methodology. By adapting the argument of Siegmund (1989), we can show that the set of all $v$ such that $Z_v$ is sufficiently close to $Z^* = \max Z_t$ is a confidence region. More precisely, we show in Proposition 3 in Appendix A that an approximate $1 - \gamma$ confidence region for $r$ is the set of all values $v$ such that

$$2Z^*Z_v^{-1}\exp[-(Z^{*2} - Z_v^2)/2\sigma^2] \geqslant \gamma \, .$$

For example, if $\sigma^{-1}Z^* = 4.04$, a 0.95 confidence region consists of all values $v$ for which $\sigma^{-1}Z_v \geqslant 2.88$. If $\sigma^{-1}Z^* = 5.68$, a 0.95 confidence region consists of those $v$ for which $\sigma^{-1}Z_v \geqslant 4.96$.

In attempting to assess the importance of a particular locus $r$ in contributing to a trait, it may be useful to estimate $\alpha$, the percentage of cases associated with that locus. In view of equation (2) we can approach this problem by estimating the parameter $\xi$. The method discussed above can be adapted to give joint confidence regions for $r$ and $\xi$ (see Siegmund 1989), but caution is necessary in interpreting the parameter $\alpha$. The method can also be adapted to give confidence regions for quantitative-trait loci in the experimental setting of Lander and Botstein (1989).

This method resembles the use of a 1- or 2-LOD support region to give a range of reasonable estimates for the recombination fraction between a trait locus and a marker. However, because $r$ is a change point, the customary explanation of that concept (Ott 1991, p. 67) is not appropriate in the present context.

### 4. *Polygenic Traits*

In general there may be more than one locus $r$ that contributes susceptibility to a trait, perhaps because of locus heterogeneity (Ott 1991, p. 198) or because particular alleles at several loci are required before an individual becomes predisposed to a trait. The tests discussed in the present paper will still be useful, but, because the value of $\alpha$ associated with any one locus is likely to be small, the process $Z_t$ may suggest regions of enriched identity by descent at several loci but may fail to exceed the threshold $b$ unless the sample size is large. A statistic designed to detect the effects of multiple loci presumably will do better, but an appropriate statistic will depend on the nature of the interaction among genes at the various loci. For example, in searching for quantitative-trait loci in experimental genetics by using the simplest Lander and Botstein (1989) model of additive effects with no interaction, in order to detect two loci contributing approximately equally to a trait, an appropriate statistic would look for a large average value, $(Z_s + Z_t)/2$, as $s$ and $t$ range over the possible loci. If the loci $s$ and $t$ under consideration are themselves linked, then the average should be divided by $[1 + \exp(-\beta|t - s|)]^{1/2}$, to account for the correlation between their genotypes. For a trait described by Risch's (1990*a*) additive model, which, as Risch notes, is approximately a model for heterogeneity, in the simple case that only two loci are involved and have approximately equal effects, the same statistic is appropriate. Problems in using Risch's (1990*a*) multiplicative model are similar. In all cases the test involves simultaneous consideration of pairs (or more) of putative loci. Determination of thresholds that account appropriately for the multiplicity of comparisons is a more complex but still tractable problem. We hope to discuss elsewhere, in greater detail, statistical analysis of problems explicitly involving multiple loci, with particular attention to the models of Risch (1990*a*, 1990*b*).

## Gaussian Approximation: More Complex Cases

The same problems arise when we consider other relative pairs and combinations of different kinds of relatives. To illustrate the situation, we first list a number of relative pairs that can be treated by minor modifications of the results already given. We then describe in somewhat more detail the particularly interesting cases of sibling pairs and sibling triples, and finally we give a brief discussion of the problem of combining different kinds of relatives.

### 1. *Other Relative Pairs*

Half-sibling, first-cousin, and aunt-niece pairs can all be treated by relatively minor modifications of the methods developed above. Feingold (in press) gives derivations of related Markov-chain models under the hypothesis of no linkage. By a variant of the reasoning given above or the calculations of Risch (1990*b*), we find that in all cases the mean value of the approximating Gaussian process $Z_t$ given by statistic (3) is of the form $\xi R(t-r)$, where $\xi = N^{1/2}\alpha p$, $p$ is the probability of identity by descent at an arbitrary locus, $\sigma^2 = p(1-p)$, and the covariance of $Z_t$ and $Z_s$ is $\sigma^2 R(t-s)$. The parameters are given below with Risch's interpretation of $\alpha$.

$$\text{half-siblings: } p = 1/2,\ \alpha = (\lambda_O - 1)/(\lambda_O + 1),$$
$$R(t) = \exp(-4\lambda|t|)\ ;$$

$$\text{aunt-niece: } p = 1/2,\ \alpha = (\lambda_O - 1)/(\lambda_O + 1),$$
$$R(t) = [\exp(-4\lambda|t|) + \exp(-6\lambda|t|)]/2\ ;$$

$$\text{first cousins: } p = 1/4,\ \alpha = 3(\lambda_o - 1)/(\lambda_o + 3),$$
$$R(t) = \exp(-4\lambda|t|)/2 + \exp(-6\lambda|t|)/3$$
$$+ \exp(-8\lambda|t|)/6\ .$$

For each of these cases the likelihood function is obtained in Proposition 1 of Appendix A. The likelihood-ratio statistic for detecting linkage is again of the form max $Z_t/\sigma$. Approximation (6) continues to hold, provided that we use for $\beta$ the weighted average of the exponents in the covariance function: $4\lambda$ for half-siblings, $5\lambda$ for aunt-niece pairs, and $16\lambda/3$ for first cousins. Approximation (7) is unchanged. Proposition 4 of Appendix A is concerned with the appropriate mathematical theory.

The essential difference among grandparent-grandchild, half-sibling, and aunt-niece pairs is in the rate at which crossovers occur—and hence in the applicable value of $\beta$. For larger values of $\beta$, slightly larger values of $b$ are required to maintain a fixed false-positive rate (see approximation [6]). For example, for our unicorn and aunt-niece pairs, for which $\beta = 0.05$, a value of $b = 4.08$ gives a false-positive rate of .05, compared with $b = 3.84$ for grandparent-grandchild pairs. As a result the approximately 90 grandparent-grandchild pairs required to achieve power of about .90 at $\alpha = 0.5(\lambda_O = 3)$ has increased to about 100 if our data come solely from aunt-niece pairs. Changes in the rate of crossovers change the effective length of the genome, as indicated in approximation (6). However, large changes in the value of $\beta$ can be compensated by relatively small changes in the value of $b$. Consequently the power—or, equivalently, the required sample size—changes surprisingly little (see fig. 1).

Cousin pairs, on the other hand, have a different value of $\alpha$—and hence of the noncentrality parameter $\xi/\sigma$. Some calculation shows that cousin pairs are less powerful than the others when $\lambda_O$ is small and that they are more powerful when it is large. The transition occurs approximately where the noncentrality parameters are equal, at $\lambda_O = 3^{1/2}$. For example, for $\lambda_O = 3$, about $N = 76$ pairs of first cousins are required to achieve 90% power, so in this case cousins are slightly more efficient than grandparent-grandchild pairs.

*Remark.* For studying homozygosity by descent along the lines suggested in the Introduction, the Gaussian process relevant to parent-child matings has the parameters $p = 1/4$, $\xi = N^{1/2}\alpha p$, $\sigma^2 = 3/16$, and $R(t) = [\exp(-2\lambda t) + \exp(-4\lambda t) + \exp(-6\lambda t)]/3$ and the same relation $EZ_t = \xi R(t-r)$ between the mean value and covariance functions as do the relative pairs discussed above. Thus the methods developed here are immediately applicable, although it presumably will rarely be the case that the sample size is large enough to make the Gaussian approximation a good one. Feingold (1993) discusses methods that are appropriate for small sample sizes.

## 2. *Siblings*

Among various family relationships for which linkage analysis using data of identity by descent of affected relative pairs is interesting, a feature peculiar (in an outbred population) to the analysis of pairs of siblings (and double cousins) is that they can be identical by descent on zero, one, or two chromosomes. If we use ordinary genetic markers—e.g., RFLPs—it is straightforward to distinguish among these three possibilities. However, in the experimental situation that we envision (Nelson et al., in press), regions of identity by descent are determined in segments without typing of individual markers within segments. The experiment to obtain these segments will be substantially simpler if one does not try to separate maternally and paternally derived chromosomes in order to distinguish whether there is identity by descent on one or on two chromosomes. Hence we consider possible analyses with and without this distinction and study the loss of power that failure to make this distinction involves.

We assume an outbred population and a trait for which the genetic contribution either shows little or no dominance effect or is sufficiently rare that its genotype is unlikely to be homozygous. For a discussion of the validity of this assumption and its role in simplifying an otherwise more complex situation, see the work of Risch (1990*b*). The information potentially available to us in $N$ sib pairs is a vector process $X_t = (X_{0,t}, X_{1,t}, X_{2,t})$, where $X_{k,t}$ denotes the number of pairs having identity by descent on $k$ chromosomes at the locus $t$. Obviously $X_{0t} + X_{1,t} + X_{2,t} = N$ for all $t$, so we can ignore any one of the coordinates. On a chromosome unlinked to a trait locus, $EX_t = N(1/4, 1/2, 1/4)$. On a chromosome containing a locus $r$ conferring susceptibility to the trait, we have $EX_r = N[(1 - \alpha)/4, 1/2, (1 + \alpha)/4]$, and $EX_t$ is easily calculated for general $t$. According to Risch (1990*b*), $\alpha = (\lambda_O - 1)/\lambda_O$. We now suppose that $N$ is large and consider the vector process $Z_t = (Z_{1,t}, Z_{2,t})$, where $Z_{1,t}$ is the large-sample Gaussian limit of $N^{-1/2}(X_{1,t} - N/2)$ while $Z_{2,t}$ is the limit of $N^{-1/2}(X_{2,t} - N/4)$. We are particularly interested in the situation where we only get to observe the sum, $Z_t^* = Z_{1,t} + Z_{2,t}$, which does not distinguish whether there is identity by descent on one or on both chromosomes.

As indicated above, the analysis of half-sibling pairs is almost identical to that given earlier for grandparent-grandchild pairs, although two meioses are involved and hence the value of $\beta$ is equal to $4\lambda = 0.04$. The basic properties of the vector process describing pairs of siblings are most easily derived by recognizing that the vector process for each sibling pair can be regarded as a sum of two independent half-sibling processes (representing the maternal and paternal meioses). From this representation, it is straightforward to compute the mean and covariance functions of the vector process $(Z_{1,t}, Z_{2,t})$. From that mean and covariance function one can obtain the following convenient alternative representation:

$$Z_{1,t} = 2U_{1,t} \qquad Z_{2,t} = U_{2,t} - U_{1,t}\,, \tag{8}$$

where $U_{1,t}$ and $U_{2,t}$ are independent Gauss-Markov processes. The covariance function of $U_{i,t}$ is $\sigma_i^2\exp(-\beta_i|t|)$ for $i = 1, 2$, where $\sigma_1^2 = 1/16$, $\beta_1 = 8\lambda$, $\sigma_2^2 = 1/8$, $\beta_2 = 4\lambda$, and $\lambda = 0.01$. The mean value of $U_{1,t}$ is always 0. On a chromosome containing no trait locus, the mean value of $U_{2,t}$ is 0; on a chromosome containing one trait locus at $r$, its mean value is $\xi \exp(-\beta_1|t - r|)$, where $\xi$ is related to the parameter $\alpha$ by the equation (see eq. [2])

$$\xi = \alpha N^{1/2}/4\,. \tag{9}$$

If experimental conditions permit observation of the vector process, $(Z_{1,t}, Z_{2,t})$, we see from relations (8) that $U_{2,t} = Z_{2,t} + Z_{1,t}/2$, and by Proposition 1 in Appendix A the log-likelihood function is $\xi(Z_{2,r} + Z_{1,r}/2)/\sigma_2^2 - \xi^2/2\sigma_2^2$. Hence the likelihood-ratio statistic to test for the presence of a trait locus is

$$\max_t [Z_{2,t} + Z_{1,t}/2]/\sigma_2\,. \tag{10}$$

Note that among all linear combinations this particular combination of $Z_{1,r}$ and $Z_{2,r}$ maximizes the noncentrality parameter, i.e., the ratio of the expectation to the SD.

If we observe only $Z_t^*$, the sum of the two coordinates of the vector process, a plausible ad hoc test statistic is

$$\max_t Z_t^*/[\sigma_1^2 + \sigma_2^2]^{1/2}\,, \tag{11}$$

but the process does not satisfy the conditions of Proposition 1 of Appendix A, so statistic (11) is not the likelihood-ratio statistic, which appears difficult to evaluate and unlikely to be substantially more powerful.

Answers to the first two of the general questions posed in the Introduction are easily obtained, provided that we use appropriate versions of approximations (6) and (7). In the case that we observe the vector process, we can immediately apply the results already derived. In comparison with grandparent-grandchild, aunt-niece, and half-sibling pairs, if we are able to use statistic (10), then sibling pairs are more informative for small $\lambda_O$ and are less so for large. The transition occurs at about $\lambda_O = (2^{1/2} - 1)^{-1}$, where the noncentrality parameters are equal. We can also find confidence regions for $r$.

Since the process $Z_t^*$ does not satisfy the conditions of Proposition 1 in Appendix A, its properties differ somewhat from those of the other processes that we have studied. Its covariance function equals $\sigma_1^2\exp(-\beta_1|t|) + \sigma_2^2\exp(-\beta_2|t|)$, which has an expansion, near 0, of the form

$$\sigma^2[1 - \beta|t| + o(|t|)]\,, \tag{12}$$

where $\sigma^2 = \sigma_1^2 + \sigma_2^2 = 3/16$ and $\beta = (\sigma_1^2\beta_1 + \sigma_2^2\beta_2)/(\sigma_1^2 + \sigma_2^2) = 0.05333\ldots$ It follows that approximation (6) and a modified form of approximation (7) apply (see Appendix A, especially approximation [A8]).

An interesting new question concerns the amount of information lost if we observe the process $Z_t^*$ instead of the vector process. This question is of direct practical importance in deciding how much experimental effort should go into obtaining data for the vector process.

Some calculations indicate that, if we can actually use statistic (10), we can do so with about one-third fewer observations than we need if we use statistic (11). A crude analysis goes as follows: The differences in behavior of the statistics are twofold. They involve different values of the parameter $\beta$ to be used in equation (6) and hence different thresholds $b$ for a given false-positive rate, say .05. As above, small changes in $b$ can accommodate rather large changes in $\beta$, so this difference plays an almost insignificant role. (For statistic [10], $\beta = 0.04$, $b = 4.02$; for statistic [11], $\beta = 0.05333 \ldots$, $b = 4.10$.) The value of $\xi$ is the same for statistics (10) and (11). The major difference is that the effective value of $\sigma^2$ to be used in the more important first term in approximation (7) and in approximation (A8) of Appendix A is much smaller when we use statistic (10) in preference to statistic (11). Since $\text{var}(Z_{2,t} + Z_{1,t}/2) = \sigma_2^2$ and $\text{var}(Z_t^*) = \sigma_1^2 + \sigma_2^2$ are the effective values of $\sigma^2$, we see from approximation (7) and equation (9) that a rough measure of relative efficiency is the ratio of the square of the noncentrality parameters, or, equivalently, $\sigma_2^2/(\sigma_1^2 + \sigma_2^2) = 2/3$. This can be translated via equation (9) into a roughly one-third-smaller sample size required to obtain a desired power if we use statistic (10) rather than statistic (11). More precise calculations using approximation (7) and approximation (A8) of Appendix A substantiate this conclusion.

Some numerical examples based on the results obtained so far are displayed in figures 1 and 2. Figure 1 gives sample sizes to achieve 90% power by using different classes of relative pairs, as a function of $\lambda_O$. Figure 2 gives the power achieved with $N$ sibling pairs, as a function of $\lambda_O$, for the values $N = 250$ and $N = 100$. To obtain the values in figure 1, say, we (i) use approximation (6) to determine the threshold $b$, then (ii) use approximation (7) (approximation [A8] when appropriate) to determine the $\xi/\sigma$ value giving the desired power, and finally (iii) use equation (2) (eq. [9] for siblings), the appropriate relation for $\alpha$ as a function of $\lambda_O$, and the appropriate value of $\sigma$ to find $N$. Once step (ii) is completed, $N$ can be expressed as a simple explicit function of $\lambda_O$.

It is readily seen that grandparent-grandchild, aunt-niece, and half-sib pairs are about equally powerful. Sibling pairs when paternal and maternal meioses are tracked are considerably more powerful than sibling pairs when meioses are not tracked separately. Siblings are relatively more powerful for small values of $\lambda_O$, while first cousins are more powerful for large values. For polygenic traits inherited according to Risch's (1990*a*) multiplicative model, the same results are valid at each individual locus, provided that the overall relative risk $\lambda_O$ is replaced by the relative-risk factor for that locus. The general situation is not so easily summarized for Risch's additive model. In the special case that the different loci make equal contributions to trait susceptibility, the effective value of $\alpha$ associated with each locus is the value corresponding to a monogenic trait, divided by the number of loci.

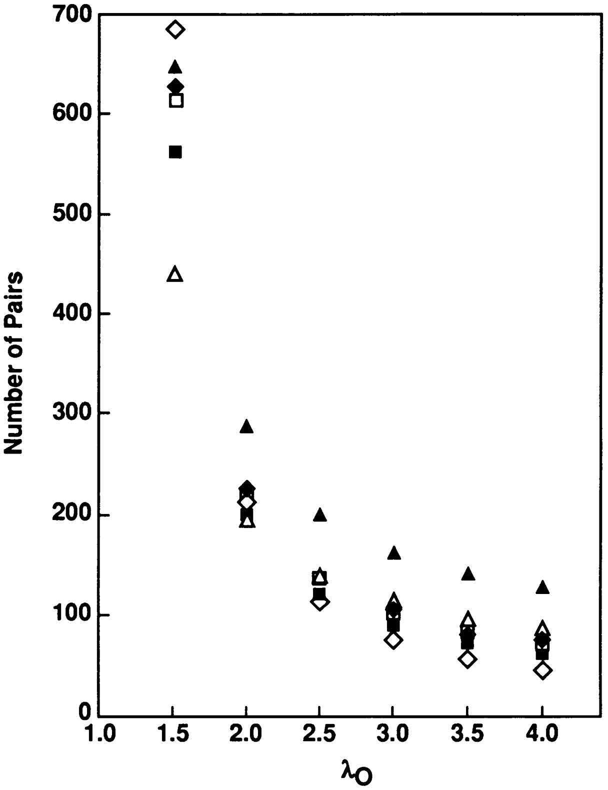

**Figure 1** Sample sizes for 90% power for different affected relative pairs, as a function of $\lambda_o$. ■ = grandparent-grandchild; □ = half-siblings; ♦ = aunt-niece; △ = siblings (tracking paternal and maternal meioses); ▲ = siblings (not tracking meioses); and ◇ = first cousins.

### 3. *Sibling Triples*

In many cases pedigrees of affected relatives will consist only of pairs, but in other cases they will consist of more than two family members. An interesting illustration is sibling triples, which we now consider.

It is useful to begin by observing that, as for sibling pairs, the identity-by-descent configuration of three siblings can be described in terms of two half-sibling configurations: that derived from maternal meioses and that derived from paternal meioses. Of the three possi-

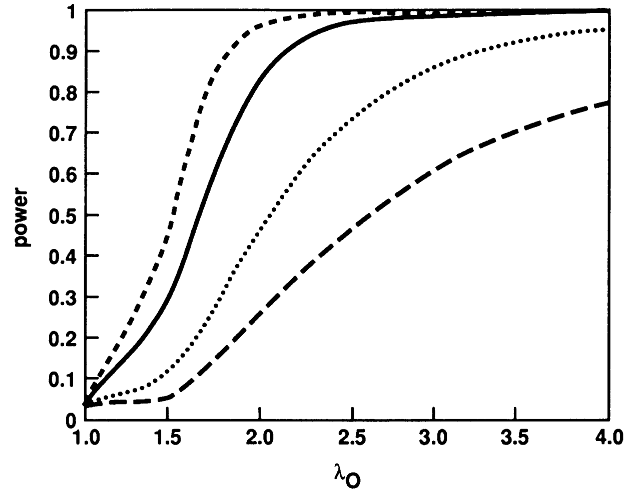

**Figure 2** Power to detect linkage by using $N$ affected sib pairs, as a function of $\lambda_o$. The shorter dashes are for $N = 250$; the dotted line is for $N = 100$, tracking maternal and paternal meioses; the solid line is for $N = 250$; and the longer dashes are for $N = 100$, not tracking meioses.

ble pairwise comparisons of three half-siblings, there can be identity by descent in exactly one or in all three comparisons, but it is impossible to have identity by descent in exactly two or in none of the comparisons. In effect the situation is exactly as it would be if we had two independent half-sib pairs, since any two comparisons determine the third. For an arbitrary locus unlinked to the trait locus, the probability of finding identity by descent in exactly one comparison is 3/4, while it is 1/4 for all three (i.e., two independent) comparisons.

A single half-sibling triple switches between the states 1 and 3. It switches from 1 to 3 at rate $\lambda$ and from 3 back to 1 at rate $3\lambda$. Let $p_{ij}(t)(i,j = 1,3)$ denote the conditional probability that the process is in state $j$ given that it is in state $i$ at a locus located a genetic distance of $t$ cM away. Standard calculations show that

$$p_{11}(t) = [3 + \exp(-4\lambda t)]/4,$$

$$p_{33}(t) = [1 + 3\exp(-4\lambda t)]/4 .$$

If we assume that we can track maternal and paternal meioses, a convenient description of the situation for full siblings is given by a pair $(i,j)$, where $i = 1$ or 3 indicates the number of pairwise comparisons showing identity by descent on the maternal chromosome and where $j$ indicates the number of pairwise comparisons showing identity by descent on the paternal chromosome. By symmetry we can combine states (1,3) and (3,1). Then we can conveniently index the states by the average $(i + j)/2$, which takes on the possible values of 1, 2, or 3. At first it appears that we may want to distinguish between those (1,1) pairs where one pairwise comparison shows identity by descent on both chromosomes and those where two different pairwise comparisons show identity by descent on exactly one chromosome. However, a more elaborate analysis in which we distinguish these possibilities does not seem to provide any useful information.

Under the hypothesis of no linkage, it is easy to write down the infinitesimal matrix for the Markov chain, $X_t = (X_{1,t}, X_{2,t}, X_{3,t})$, where the $k$th coordinate denotes the number of sibling triples indexed by $(i + j)/2 = k$. The stationary distribution is $N(9/16, 3/8, 1/16)$. Since the sum of the three coordinates equals the total number, $N$, of triples, it suffices to consider only two coordinates, say the second and third. The approximating Gaussian process, $Z_t = (Z_{2,t}, Z_{3,t})$ is obtained as an approximation to the centered and scaled process $(X_{2,t} - 3N/8, X_{3,t} - N/16)/N^{1/2}$. From the representation in terms of independent half-sibling processes and the conditional probabilities given above, it is straightforward to calculate the covariance function of the process $Z_t$, which is given below in relations (14) and relations (16). If we are unable to track maternal and paternal meioses, instead of $X_t$ we can observe only the sum $X_t^* = X_{2,t} + X_{3,t}$, which counts the total number of triples for which the three possible pairwise comparisons show identity by descent on at least one chromosome. The associated Gaussian process is $Z_t^* = Z_{2,t} + Z_{3,t}$.

On a chromosome containing a locus $r$ at which some allele contributes to the trait of interest, it may be shown that $EX_r$ is of the form $N[9(1 - \alpha)/16, 3(1 + \alpha)/8, (1 + 3\alpha)/16]$ for some $0 < \alpha < 1$. A derivation of this expression, together with an expression for $\alpha$ in terms of Hardy-Weinberg frequencies and their penetrances, can be obtained along the lines of an argument by Risch (1990*b*), but, since the calculation is somewhat complicated and requires the introduction of some new parameters, we defer it to Appendix B. It follows from the Haldane mapping function that, along the chromosome containing the locus $r$,

$$(EZ_{2,t}, EZ_{3,t}) = \xi \exp(-\beta|t - r|)(1/2, 1/4) , \quad (13)$$

where $\xi = 3\alpha N^{1/2}/4$, $\beta = 4\lambda = 0.04$.

As in the case of sibling pairs, it is convenient to represent the process $Z_t$ in terms of independent processes $U_{2,t}$, $U_{3,t}$. Appropriate processes can be inferred from the mean value and covariance function of $Z_t$ and are defined through the relations

$$\begin{aligned} Z_{2,t} &= (U_{2,t} + U_{3,t})/2, \\ Z_{3,t} &= (U_{3,t} - U_{2,t})/4 . \end{aligned} \tag{14}$$

The mean values and covariance functions of the $U$'s are given by

$$EU_{2,t} = 0, \qquad EU_{3,t} = \xi \exp(-\beta|t - r|) , \tag{15}$$

and

$$\begin{aligned} \operatorname{cov}(U_{2,t}U_{2,s}) &= \sigma_2^2\exp(-2\beta|t - s|), \\ \operatorname{cov}(U_{3,t}U_{3,s}) &= \sigma_3^2\exp(-\beta|t - s|) , \end{aligned} \tag{16}$$

where $\sigma_2^2 = 9/16$, $\sigma_3^2 = 3/8$.

In the case that we are able to track maternal and paternal meioses, Proposition 1 in Appendix A allows us to evaluate the likelihood function and show that the likelihood-ratio statistic is

$$\max U_{3,t}/\sigma_3 = \max(Z_{2,t} + 2Z_{3,t})/\sigma_3 . \tag{17}$$

If we are experimentally unable to track maternal and paternal meioses and are limited to observations on $Z_t^* = Z_{2,t} + Z_{3,t} = (U_{2,t} + 3U_{3,t})/4$, a plausible test statistic is

$$\max Z_t^*/(\sigma_2^2/16 + 9\sigma_3^2/16)^{1/2} . \tag{18}$$

We can use the techniques developed previously to analyze these statistics, and we find that statistic (18) is approximately 6/7 as efficient as statistic (17). This is a considerably more favorable ratio than the 2/3 we obtained in the case of sibling pairs.

## 4. *Combining Different Classes of Relatives*

An important problem is to combine efficiently data from different classes of relatives. A very simple version of this question arises in the study of sib pairs when one asks how the coordinates of the vector $Z_t = (Z_{1,t}, Z_{2,t})$ should be combined to get an overall test statistic. The problem seems reasonably tractable when different relative pairs are independent. We can calculate, at least approximately, a joint likelihood function. It will in general involve different values of the parameter $\alpha$ for the different relative pairs. Although these parameters are unknown, in the models of Risch (1990*a,* 1990*b*) they have specific relations to each other. In some cases we are able to use these relations to show that the likelihood-ratio test statistic involves the maximum of a new process, which is a particular linear combination of the corresponding processes for the different classes of relative pairs.

To consider the general question of combining data from two different classes of relatives, suppose that we have two independent processes, say $W_{i,t}(i = 1,2)$. For example, $W_{1,t}$ and $W_{2,t}$ might be the processes in the numerators of statistics (10) and (17) or in statistics (11) and (18). Assume that at the trait locus $r$

$$EW_{1,r} = \mu N_1^{1/2}, \qquad EW_{2,r} = c\mu N_2^{1/2} , \tag{19}$$

where the $N_i$ are the sample sizes on which the two processes are based, while $c$ and $\mu$ are positive parameters. Assume also that $\operatorname{var}(W_{i,t}) = v_i^2$.

The parameter $\mu$ is usually unknown. In some cases it may be reasonable to regard $c$ as known, at least approximately. An example would be grandparent-grandchild and aunt-niece pairs, where, for Risch's (1990*a,* 1990*b*) single-locus, additive-multilocus, and multiplicative-multilocus models, the value of $c$ is 1. For other combinations of relative pairs, it may be more difficult or impossible to establish a reasonable value for $c$, especially if that value would depend on the mode of inheritance of the trait. Usually we shall want to combine several classes of relatives, with some combinations falling into each of these cases. We begin with the simpler case, where $c$ can be assumed known.

If the covariance function of each of the processes $W_{i,t}$ has the structure assumed in Proposition 1 of Appendix A (statistics [10] and [17] do, but statistics [11] and [18] do not), we obtain from that proposition a simple expression for the likelihood function of the combined data. From this likelihood function one easily sees that the likelihood-ratio statistic for testing the hypothesis that $\mu = 0$ is of the form $\max W_t/v$, where

$$W_t = N_1^{1/2}W_{1,t}/v_1^2 + cN_2^{1/2}W_{2,t}/v_2^2 \tag{20}$$

and

$$v^2 = N_1/v_1^2 + c^2N_2/v_2^2 .$$

The theory that we have developed can be applied to the process $W_t$, to obtain approximations to the significance level and power of the test and approximate confidence regions for $r$. If the means and covariances of the $W_{i,t}$ do not have the structure of Proposition 1 in Appendix A, then it still seems reasonable to use the linear combination (20), which maximizes the noncentrality parameter at $t = r$.

A simple example is grandparent-grandchild and aunt-niece pairs. If we are willing to assume, on the basis of Risch's (1990*a*, 1990*b*) models, perhaps supplemented by epidemiological evidence, that $c = 1$, then the preceding argument shows that the likelihood-ratio test combines aunt-niece and grandparent-grandchild pairs, using equal weights. Actually this combination is not exactly optimal. Because of their higher rate of recombination, aunt-niece pairs require a larger value of the threshold $b$ than do grandparent-grandchild pairs. Consequently aunt-niece pairs are less efficient and should receive less weight in the combined data. As indicated above, however, this effect is slight, and equal weights are satisfactory for practical purposes. If we are limited in the determination of regions of identity by descent by reference to a fixed discrete set of markers, the different recombination rates play a more important role, so the situation must be examined more closely (see Bishop and Williamson 1990; Risch 1990*b*).

*Remarks.* (i) A similar situation exists with respect to combining half-siblings with grandparent-grandchild pairs, aunt-niece pairs, or both. (ii) A straightforward calculation using the representations of Feingold (in press) shows that, even when the niece is the same person as the grandchild, the two possible pairwise comparisons are stochastically independent. Consequently, unlike the case of sibling triples, it is unnecessary to consider such a triple as being a new pedigree; we can amalgamate the grandparent-grandchild comparison with others of that kind, and we can do likewise with the aunt-niece comparison. Although this procedure would not be an optimal use of these pedigrees, some calculation shows that it is reasonable. (iii) The preceding analysis shows that it is easy to incorporate different crossover rates for male and female meioses. For each class of relatives, we create subclasses corresponding to the meioses involved (e.g., we divide grandparent-grandchild pairs into two groups according to whether the meiosis involves the grandchild's mother or father). We then use the theory given above, to combine the subclasses. This results in a class of relative pairs with a value of $\beta$ averaged according to the proportions of the different meioses. Simple calculations show that such an analysis is unlikely to have a substantial impact. The situation would be more complicated and require more careful analysis if we try to account for heterogeneous variation in male and female crossover rates along the genome.

For a second example, suppose that we want to combine sibling pairs and sibling triples. This application poses some difficulties, because there is no obvious relation between the parameters $\alpha$ for the two cases and hence no natural value for the constant $c$. In particular, there is no reason to assume that the two parameters are the same, although they may be in some cases. If we use one of Risch's (1990*a*, 1990*b*) models, the parameters $\alpha$ are specific functions of the Hardy-Weinberg frequencies of the various alleles at the locus $r$ and of their penetrances. One possibility is to assume hypothetical values for these parameters and to investigate the sensitivity of the resulting procedures to these assumptions.

For the model of monogenic inheritance, of Risch (1990*a*, 1990*b*), we have in Appendix B obtained expressions for $\alpha$, for both sibling pairs and triples, in terms of Hardy-Weinberg frequencies and penetrances. These values are approximately equal when only one very rare allele has positive penetrance. If the Hardy-Weinberg frequency of the one allele is in the range .1–.25, then the parameter $\alpha$ for sibling pairs is about 25%–40% larger than that for sibling triples.

If we want to combine sibling pairs with sibling triples, under the condition that we are able to track maternal and paternal meioses, and if we assume that $c = 3$, which is appropriate if the values of the parameter $\alpha$ are the same for both pairs and triples, then the weights $1/v_1^2$ and $c/v_2^2$ in equation (20) are equal. In the event that we are unable to track individual meioses, the triples entering into statistic (18) should get 12/7 times the weight of the pairs entering into statistic (11). If we assume that the value of $\alpha$ for sibling pairs is $\rho$% larger than that for sibling triples, then the relative weight for pairs should be increased by $\rho$% in each case. It would be interesting to see whether something like this analysis applies to Risch's (1990*a*, 1990*b*) polygenic models.

If it is difficult or impossible to choose a value for $c$, it can be regarded as unknown. In that case the likelihood-ratio statistic is

$$\max[(W_{1,t}^{+}/v_1)^2 + (W_{2,t}^{+}/v_2)^2]^{1/2}\,, \tag{21}$$

where $a^+$ denotes the maximum of $a$ and 0. It is possible to analyze statistic (21) and higher-dimensional generalizations for dealing with more than two classes of relatives (see approximation [A11] in Appendix A). In cases where we can make an educated guess at a value for $c$, it appears from preliminary calculations that statistic (20) will turn out to be more efficient than statistic (21), unless our guess is poor. On the other hand, the penalty for using statistic (21) or a higher-dimensional version appears to be small unless the number of classes of relatives involved is large. For example, suppose we combine approximately equal numbers of independent grandparent-grandchild and half-sibling pairs, using statistic (21) rather than the equally weighted linear combination suggested by the preceding analysis. By approximation (A11) in Appendix A, $b = 4.28$ (LOD 3.98) is the appropriate threshold for a false-positive rate of .05, and additional calculations (not shown) indicate that only about 10% more affected pairs would be required to obtain 90% power. If our relative pairs are divided approximately equally among grandparent-grandchild, half-sibling, and aunt-niece pairs, about 15% more pairs would be required.

If several classes of relatives are involved, then there will be a greater loss of efficiency if we use a several-dimensional version of statistic (21) instead of an appropriate linear combination. In practice it seems reasonable to use a combination of the two methods: linear combinations to combine classes of relatives for which one can make reasonable assumptions relating the noncentrality parameters (e.g., grandparent-grandchild, half-siblings, and aunt-niece) and a higher-dimensional version of statistic (21) to pool the remaining classes (some of which will consist of linear combinations of smaller classes) into an overall statistic. We expect to make a more detailed study of this problem in the future.

## Discrete Set of Markers

Until now we have assumed that the process $Z_t$ can be observed continuously, i.e., that for practical purposes we know exactly where regions of identity by descent begin and terminate. However, if the process is determined by reference to a discrete set of markers, say one at every $\Delta$ cM along the genome, then the appropriate theory is more difficult. (For a discussion of this issue as it relates to the search for quantitative trait loci in experimental genetics, see the paper by Lander and Botstein [1989].) As an illustration consider the special case of grandparent-grandchild pairs. A simple test statistic is the maximum value of the process $Z_t$ as $t$ runs through the equally spaced set of markers, although this is not the actual likelihood-ratio statistic, which is somewhat more complicated. Its significance level is approximately (see Proposition 4 in Appendix A)

$$1 - \Phi(b) + \beta l b \phi(b) v[b(2\beta\Delta)^{1/2}], \qquad (22)$$

where $v(x)$ is a special function which can be evaluated numerically and is reasonably well approximated by $\exp(-0.583x)$ (Siegmund 1985, chap. 4). For grandparent-grandchild pairs of the unicorn discussed above, for 1-, 5-, 10-, and 20-cM maps the 0.05 false-positive thresholds are, respectively, approximately $b = 3.72$, 3.59, 3.50, and 3.38 (LOD 3.01, 2.80, 2.66, and 2.48, respectively). In effect, if we test fewer markers, the threshold $b$ can be lowered slightly while the same overall false-positive rate is maintained.

Discussions of power and confidence regions become appreciably more complex when regions of identity by descent are determined with reference to a discrete set of markers. Specifically, the power of the test suggested in the preceding paragraph exhibits greater variability as a function of both the location of the trait locus with respect to the markers and the relationship of the affected pairs under consideration. If the trait locus is sufficiently close to a marker, then the lower threshold means that the power can actually increase slightly. However, if it is located approximately midway between markers, then the power can decrease appreciably. For more distant relatives, whose identity-by-descent regions involve several meioses, there is a greater chance for multiple crossovers to occur between markers than there is in, say, grandparent-grandchild pairs, which involve only a single meiosis. Hence the test has comparatively less power to detect a trait locus lying midway between markers, when more distant relatives predominate. This problem can be alleviated somewhat by using the true likelihood-ratio statistic, which is related to Lander and Botstein's (1989) "interval mapping." If $\Delta$ is small enough in relation to the composition of the relatives under consideration, then a discrete set of markers will behave more or less like a continuous one. A basic question is to determine how small this $\Delta$ must be. These issues will be discussed in detail elsewhere.

Approximation (22) is also appropriate in the context of Lander and Botstein's (1989) work. It can be seen to be quite accurate by comparison with the simulations presented there.

## Discussion

In this paper we have developed approximate Gaussian models for statistical analysis of genetic linkage, using complete high-resolution maps of identity by descent of affected relative pairs. We have determined thresholds to control the overall false-positive rate of tests to detect linkage and have developed approximations to the power of the tests, from which one can infer the sample size necessary to detect an hypothesized effect with a specified probability. We have also discussed confidence regions for a trait locus.

There are some basic differences between our approach and that of testing linkage to individual markers, as exemplified by Risch (1990*b*). Risch uses the traditional LOD-3 threshold for proving linkage, without regard to either the number of markers tested or the composition of the relative pairs studied. This criterion would not be sufficiently stringent and would lead to unacceptably high false-positive rates if it were used to test a continuous map of markers. Moreover, the actual false-positive rate would vary from one class of relatives to another, making comparisons of power potentially misleading. Our method uses a threshold that is appropriate to both the composition of the relative pairs under consideration and the test statistic employed. The statistical price of using a continuous map and, hence, a more stringent threshold is that our test is slightly less powerful than Risch's in the case that one of his markers is at zero recombination distance from a trait locus. Usually this will not be the case when linkage to individual markers is tested, and then our test can be considerably more powerful.

We have also introduced some ideas for efficiently combining data from different classes of relatives, and we expect to examine this issue in more detail in the future. In addition, we hope to develop specific statistical procedures for models dealing with complex modes of inheritance and to compare their performance with the performance of the methods developed in this paper.

## Acknowledgments

The research of D.S. has been supported in part by NSF grant DMS91-04432 and by NSA grant MDA 909-92-H-3092. The research of P.O.B. is supported by the Howard Hughes Medical Institute.

## Appendix A

A family of random variables $\{Z_t, -\infty < t < \infty\}$ is called a "Gaussian process" if, for each $n = 1, 2, \ldots$ and $t_1 < \ldots < t_n$, the random variables $Z_{t_1}, \ldots, Z_{t_n}$ are jointly normally distributed. A Gaussian process is specified by its mean value or drift function $\mu(t) = \mathrm{E}(Z_t)$ and covariance function $\mathrm{C}(s,t) = \mathrm{Cov}(Z_s, Z_t)$. The process is said to be covariance stationary if $\mathrm{C}(s,t)$ is a function only of $t - s$, say $\mathrm{C}(s,t) = \sigma^2 R(t - s)$.

The value $r$ is called a "change point" (for $\mu[t]$) if there is a jump in the derivative $\mu'(t)$ at $t = r$. A common example is Brownian motion with a broken-line drift, for which $\mathrm{C}(s,t) = \min(s,t)$ for $s,t \geqslant 0$ and $\mu(t;r) = \mu_0 t + \delta(t - r)^+$. Here $a^+ = \max(a,0)$. Most of the examples of this paper are covariance stationary and have a mean value function of the form $\mu(t;r) = \xi R(t - r)$ with a change point at $r$. The special case that $R(t) = \exp(-\beta|t|)$ and $\xi = 0$ is called the "Ornstein-Uhlenbeck process."

In the following discussion there are some unstated technical regularity conditions on the function $R(t)$, which are trivially satisfied in all cases of interest, where $R(t)$ is a finite linear combination of exponentials with positive coefficients (see Leadbetter et al. 1983, chap. 12). Some of the mathematical details will be discussed elsewhere.

*Proposition 1.* Let $\{Z_t, -\infty < t < \infty\}$ be a stationary Gaussian process with mean 0 and covariance function

$$\mathrm{Cov}(X_s, X_t) = \sigma^2 R(t - s)\,, \tag{A1}$$

where $R(0) = 1$. Assume

$$\int_{-\infty}^{\infty} |R(t)|\,dt < \infty\,.$$

Let $\xi$, $r$ be arbitrary and let

$$\mu(t;r) = \xi R(t - r)\,. \tag{A2}$$

Let $P_1$ denote the distribution of $\{Z_t + \mu(t;r), -\infty < t < \infty\}$. Then the likelihood function is

$$\frac{dP_1}{dP}(Z) = \exp(\xi Z_r/\sigma^2 - \xi^2/2\sigma^2)\,.$$

*Proof.* The proof is a straightforward application of the representation given by Parzen (1961, theorem 7A) or can be derived from first principles by passing to the limit from the finite-dimensional case.

*Lemma 1.* Let $Z_t$ be a stationary Gaussian process process with mean 0 and covariance function (A1). Assume that as $t \rightarrow 0$,

$$R(t) = 1 - \beta|t| + o(|t|)\,. \tag{A3}$$

Let $0 < x < b$ and define $t^*$ to be the unique positive solution of the equation $R(t^*) = xb^{-1}$. Let $t_1 > 0$. Let $b$

and $x$ be large, and assume that $t^*$ is contained in $(0,t_1)$ and is bounded away from the upper endpoint. Then

$$P\{\max_{0<t<t_1} Z_t \geqslant b \mid Z_0 = x\} \sim \beta |R'(t^*)|^{-1}\exp[-(b^2 - x^2)/2\sigma^2]\,. \tag{A4}$$

When $x$ is close to $b$, approximation (A4) can be expressed in the form

$$P\{\max_{0<t<t_1} Z_t \geqslant b \mid Z_0 = x\} \sim \exp[-(b - x)x/\sigma^2] \tag{A5}$$

when $0 < b - x \rightarrow 0$.

*Proof.* In the special case of an Ornstein-Uhlenbeck process these results are essentially equivalent to results that Siegmund (1985, chap. 4) gives for Brownian motion. In the general case they can be derived by suitable modifications of the methods of Woodroofe (1976, 1982) and Siegmund (1985). These arguments also yield appropriately modified approximations for the case when $t$ is limited to a discrete set of values. See Proposition 4 below.

*Proposition 2.* Let $Z_t$ be a covariance-stationary Gaussian process with covariance function (A1) satisfying equation (A3) and with mean value

$$E(Z_t) = \xi R(t - r),\ -\infty < t < \infty\,. \tag{A6}$$

For any fixed $0 < r < l$ and large $b$ and $\xi$,

$$P\{\max_{0<t<l} Z_t/\sigma > b\} = 1 - \Phi(b - \xi/\sigma) + \phi(b - \xi/\sigma)[2(\xi/\sigma)^{-1} - (\xi/\sigma + b)^{-1}][1 + o(1)]\,. \tag{A7}$$

*Proof.* The argument is similar to that of the proof of equation (32) of James et al. (1987) or corollary (4.19) of Siegmund (1985), so we briefly sketch it. The first term of equation (A7) accounts for the possibility that $Z_r > b\sigma$. The argument is completed by (i) conditioning on $Z_r = x < b\sigma$, (ii) observing that, by Proposition 1, if $r$ is known, then $Z_r$ is sufficient for $\xi$, so the conditional process behaves like a mean zero process under the same conditioning, and (iii) using approximation (A5) of Lemma 1 three times: to account for the possibilities that $Z_t > b\sigma$ for some value of $t > r$, for some value of $t < r$, and for both. Technical aspects of the proof involve showing that only values of $x$ in the range where approximation (A5) is applicable make a non-negligible contribution.

*Remark.* Proposition 2 does not apply directly to the statistic $\max Z_t^*/\sigma$ of statistic (11) (i.e., $\sigma = \sigma_1^2 + \sigma_2^2)^{1/2}$, since in that case the mean and covariance function do not have the relation in covariance function (A1) and equation (A6). Nevertheless, by a slightly different argument—that, if instead of equation (A6), we have $E(Z_t) = \xi R_1(t - r)$, where $R_1(t) = 1 - \beta_1|t| + o(|t|)$, then

$$P\{\max Z_t^* > b\sigma\} \approx 1 - \Phi(b - \xi/\sigma) + \phi(b - \xi/\sigma)\{2\beta\sigma/\beta_1\xi - [\xi\sigma^{-1}(2\beta_1/\beta - 1) + b]^{-1}\}\,, \tag{A8}$$

provided that $(1 - \beta_1/\beta)\xi/(b\sigma) < 1$. This restriction on the range of $\xi$ is immaterial for most applications. The process $Z_t^*$ of statistic (11) is the special case $\beta = 16\lambda/3$, $\beta_1 = 4\lambda$, $\sigma^2 = 3/16$.

*Proposition 3.* For a covariance-stationary Gaussian process $\{Z_t\}$ satisfying covariance function (A1) and equations (A3) and (A6) with $0 < r < l$, for large values of $Z^* = \max_{0<t<l} Z_t$, an approximate $1 - \gamma$ confidence region for $r$ is the set of all $v$, $0 < v < l$ satisfying

$$2\beta|R'(t^* - v)|^{-1}\exp[-(Z^{*2} - Z_v^2)/2\sigma^2] \geqslant \gamma\,,$$

where $R(t^* - v) = Z_v/Z^*$.

*Proof.* We follow the argument of Siegmund (1989). We first consider a test of the hypothesis that $r = v$ that rejects if $Z^* - Z_v > c$, for an appropriate value of $c$. Under the hypothesis, it follows from Proposition 1 that $Z_v$ is a sufficient statistic for the nuisance parameter $\xi$, so the conditional probability $P\{Z^* - Z_v > c \mid Z_v\}$ does not depend on the unknown value of $\xi$. If we choose $c = c(Z_v)$ so that this conditional probability is equal to $\gamma$, then the set of values $v$ that are not rejected by the test gives a confidence region for $r$. If we use approximation (A4) to approximate the required probability, then the resulting approximate confidence region is the one described in the proposition.

*Proposition 4.* Let $Z_t$ be the stationary Gaussian process described in Lemma 1. Let $l > 0$. Suppose that $b \rightarrow \infty$, $\Delta \rightarrow 0$ in such a way that $b\Delta^{1/2}$ converges to a finite constant. Then

$$P\{\max_{0\leqslant i\Delta\leqslant l} \sigma^{-1}Z_{i\Delta} \geqslant b\} \sim \beta lb\phi(b)\upsilon[b(2\beta\Delta)^{1/2}]\,, \quad \text{(A9)}$$

where $\upsilon$ is a special function described by Siegmund (1985, p. 82). In the limiting case of continuous observation, the corresponding asymptotic result is

$$P\{\max_{0\leqslant t\leqslant l} \sigma^{-1}Z_t \geqslant b\} \sim \beta lb\phi(b)\,. \quad \text{(A10)}$$

*Proof.* The result follows by integrating the appropriate version of approximation (A4) of Lemma 1, which, for a discrete time process, contains the additional factor $\upsilon[b(2\beta\Delta)^{1/2}]$. In fact, a complete proof of approximation (A4), along the lines suggested in Lemma 1, begins with a derivation of the discrete-time result and requires additional technical arguments to obtain the continuous-time version as well. Alternatively one can avoid reference to approximation (A4), by applying directly the method of proof of Lemma 1 to the unconditional probability in approximation (A9).

*Remarks.* (i) The implication of the condition that $b\Delta^{1/2}$ converges to a finite limit is that the argument of the function $\upsilon$ should not be extremely large. As a practical matter, it entails that the number of markers per chromosome should not be too small. (ii) Approximations (6) and (22) contain, in addition to expressions (A9) and (A10), a boundary term to account for the possibility that $Z_0 > b\sigma$. For small values of $\Delta$—e.g., in the limiting case of continuous mapping when $\Delta = 0$—this term plays a comparatively insignificant role, but for larger values it can make an important contribution to the overall approximation. (iii) The approximation (A5) version that is suitable for a discrete-time process is

$$P\{\max_{0<i\Delta<t_1} \sigma^{-1}Z_{i\Delta} \geqslant b \,|\, \sigma^{-1}Z_0 = x\} \approx \exp[-(b-x)x]\upsilon[b(2\beta\Delta)^{1/2}]\,.$$

For discussion of a similar approximation in a related context, see the work of James et al. (1987) or Siegmund (1985, chap. 9). This approximation can be combined with the proof of Proposition 2 to give an approximation to the power of the test proposed in the Combining Different Classes of Relatives subsection, when observations are made at a discrete set of markers spaced at distances $\Delta$ along a chromosome. In the case where $r$ is at zero recombination distance from a marker, the argument is completely straightforward and leads to an approximation where the first term in square brackets in (A7) is simply multiplied by the correction factor $\upsilon[b(2\beta\Delta)^{1/2}]$, while the second term is multiplied by the square of the same correction factor. In the case that $r$ lies in the interval between two markers, the situation is slightly more complicated and involves a numerical integration. The details are omitted. (iv) A more complicated calculation along the lines of Proposition 4 gives approximations to the false-positive rate for statistic (21). If $W_{1,t}$ and $W_{2,t}$ both satisfy the conditions of Lemma 1—with the parameters $\beta_1$ and $\beta_2$, respectively—on a chromosome that is unlinked to the trait of interest, then the probability that statistic (21) exceeds $b$ approximately equals

$$2^{-1}(\beta_1 + \beta_2)lb\exp(-b^2/2)[b/4 + (2\pi)^{-1/2}]\,. \quad \text{(A11)}$$

Approximation (A11) is also appropriate in the context of the work of Lander and Botstein (1989), if one considers the progeny of an $F_1$ intercross and permits a dominance component in the model.

## Appendix B

In this appendix we generalize some of the calculations by Risch (1990*a*, 1990*b*), to family constellations consisting of three members. As in the body of the paper, we assume no dominance effect.

Let $\varphi_j$ denote the phenotype of the $j$th member of a family, i.e., $\varphi_j = 1$ or 0 according as the $j$th member is affected or not. Let $K = E\varphi_j$ be the probability of an individual's being affected. In the absence of a dominance effect we have the representation

$$\varphi_j = K + f_{x_j} + f_{y_j}\,, \quad \text{(B1)}$$

where $f_{x_j}$ ($f_{y_j}$) is the centered penetrance of the allele inherited from the $x(y)$ parent. "Centered" means that $K/2$ has been subtracted from the actual penetrance, so $Ef_{x_j} = Ef_{y_j} = 0$. See Kempthorne (1957, p. 330 ff.). Let $K_R$ be the conditional probability that a type-$R$ relative of an affected is also affected, so $KK_R = E(\varphi_1\varphi_2)$, and let $\lambda_R = K_R/K$ be the relative risk of a type-$R$ relative of an affected. By simple manipulations Risch (following James [1971]) observes that

$$\lambda_R = 1 + K^{-2}\text{cov}(\varphi_1,\varphi_2)\,,$$

and

$$\text{cov}(\varphi_1,\varphi_2) = (\textstyle\sum p_a f_a^2)\gamma_R\,. \quad \text{(B2)}$$

Here $p_a$ is the Hardy-Weinberg frequency of allele $a$ and $\gamma_R = P(x_1 \equiv x_2) + P(x_1 \equiv y_2) + P(y_1 \equiv x_2)$

$+ P(y_1 \equiv y_2)$, where $(x_1 \equiv x_2)$, etc., denote identity by descent of the indicated alleles; $\gamma_R$ is four times the coefficient of parentage (Kempthorne 1957, pp. 73–74). Some additional manipulations allow Risch to express $\lambda_R$ for various relations $R$ in terms of $\lambda_o$, the special case of parent and offspring. He then uses Bayes's formula to give an expression involving $\lambda_o$ for the conditional probability of a particular identity-by-descent relation between two relatives, given that both are affected. When we assume Risch's (1990*b*) model, the parameter $\alpha$ is a simple function of this conditional probability. In the special case of siblings, if there is only one allele of Hardy-Weinberg frequency $q$ having positive penetrance, then

$$\alpha = (\lambda_0 - 1)/\lambda_0 = (1 - q)/(1 + 3q)\,,$$

which is independent of the value of the penetrance.

To consider three relatives, we begin by evaluating the relative risk that all three are affected,

$$K^{-3}\mathrm{E}(\varphi_1\varphi_2\varphi_3) = 1 + K^{-2}\sum_{j<k} \mathrm{cov}(\varphi_j,\varphi_k) + K^{-3}\mathrm{E}[(\varphi_1 - K)(\varphi_2 - K)(\varphi_3 - K)]\,. \tag{B3}$$

Using representation (B1), we can easily evaluate the final expectation in equation (B3) as

$$[\textstyle\sum p_a f_a^3][P(x_1 \equiv x_2 \equiv x_3) + \ldots + P(y_1 \equiv y_2 \equiv y_3)]\,. \tag{B4}$$

For three siblings, the first and last probabilities appearing in formula (B4) each equal 1/4, and the other six equal 0 (because the population is assumed to be outbred).

To compute via Bayes's formula the conditional probability of an identity-by-descent configuration given three affected, we must also evaluate the covariances and final expectation in equation (B2), conditional on the identity-by-descent configuration. For three siblings, when we condition on three positive identity-by-descent comparisons for each of the two alleles, the probabilities in formula (B4) sum to 2. They sum to 1 when given three positive identity-by-descent comparisons on one allele and one positive comparison on the other. Given one positive identity-by-descent comparison on each allele, the coefficient $\gamma_R$ in equation (B2) equals 2/3; given three positive identity-by-descent comparisons on one allele and one on the other, it equals 4/3. Other coefficients can be evaluated similarly, with the result that, for sibling triples,

$$\alpha = [K^{-2}\textstyle\sum p_a f_a^2 + (2K^3)^{-1}\sum p_a f_a^3]/ [1 + 3K^{-2}\sum p_a f_a^2 + (2K^3)^{-1}\sum p_a f_a^3]\,.$$

In the special case of a single allele having positive penetrance, this becomes

$$\alpha = [1 + q - 2q^2]/[1 + 9q + 6q^2]\,.$$

## References

Aldous D (1989) Probability approximations via the Poisson clumping heuristic. Springer, New York

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

Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314–331

Feingold E. Markov processes for modeling and analyzing a new genetic mapping method. J Appl Probability (in press)

——— (1993) Modeling a new genetic mapping method. PhD thesis, Stanford University, Stanford

James B, James KL, Siegmund D (1987) Tests for a change-point. Biometrika 74:71–83

James JW (1971) Frequency in relatives for an all-or-none trait. Ann Hum Genet 35:47–48

Kempthorne O (1957) An introduction to genetic statistics. John Wiley & Sons, New York

Lander ES, Botstein D (1986*a*) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb Symp Quant Biol 51:49–62

——— (1986*b*) Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphims. Proc Natl Acad Sci USA 83:7353–7357

——— (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570

——— (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Leadbetter MR, Lindgren G, Rootzén H (1983) Extremes and related properties of random sequences and processes. Springer, New York

Nelson SF, McCusker JH, Sander MA, Kee Y, Modrich P, Brown PO. Genomic mismatch scanning: a new approach to genetic linkage mapping. Nature Genet (in press)

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

Parzen E (1961) An approach to time series analysis. Ann Math Stat 32:951–989

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

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

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

Siegmund D (1985) Sequential analysis: tests and confidence intervals. Springer, New York

——— (1989) Confidence sets in change-point problems. Int Stat Rev 56:31–48

Woodroofe M (1976) A renewal theorem for curved boundaries and moments of first passage times. Ann Probability 4:67–80

——— (1982) Nonlinear renewal theory in sequential analysis. Society for Industrial and Applied Mathematics, Philadelphia