On the power for linkage detection using a test based on scan statistics

Biostatistics (2005), **6**, 2, *pp.* 259–269
doi:10.1093/biostatistics/kxi007

On the power for linkage detection using a test based on scan statistics

SONIA HERNÁNDEZ*

Department of Statistics, Universidad Rey Juan Carlos, C/Tulipán, s/n, 28933 Móstoles, Madrid, Spain
shernandeza@escet.urjc.es

DAVID O. SIEGMUND

Department of Statistics, Serra Mall, Sequoia Hall, Stanford University, Stanford, CA, USA

MATHISCA DE GUNST

Department of Mathematics, Vrije Universiteit, Amsterdam, and EURANDOM, PO Box 513-5600 MB Eindhoven, The Netherlands

SUMMARY

We analyze some aspects of scan statistics, which have been proposed to help for the detection of weak signals in genetic linkage analysis. We derive approximate expressions for the power of a test based on moving averages of the identity by descent allele sharing proportions for pairs of relatives at several contiguous markers. We confirm these approximate formulae by simulation. The results show that when there is a single trait-locus on a chromosome, the test based on the scan statistic is slightly less powerful than that based on the customary allele sharing statistic. On the other hand, if two genes having a moderate effect on a trait lie close to each other on the same chromosome, scan statistics improve power to detect linkage.

Keywords: Complex traits; Gaussian processes; Genetic linkage; Scan statistics; Score test.

1. INTRODUCTION

Motivated by Terwilliger *et al.* (1997), who claimed that "true peaks" in the sample paths of allele sharing statistics were wider than "false peaks," and that this information might be used profitably in linkage analysis, Hoh and Ott (2000) suggest that the power to detect genetic linkage could be improved by combining the information on several contiguous markers. In particular, they propose to use a moving sum (or equivalently a moving average) of the values at several consecutive markers of a statistic instead of the statistic itself. Such a statistic might be, for example, the proportion of alleles shared identical-by-descent (IBD) for pairs of relatives or the logarithm of the likelihood ratio (usually called logarithm of odds, or LOD score, by biologists) in pedigrees. Hoh and Ott call these moving averages *scan statistics* and compute the corresponding p-values by Monte Carlo permutation tests. Although they do not make a thorough

*To whom correspondence should be addressed.

study of these statistics, in an application to autism families they find a region that was missed with the standard approach. The value of scan statistics has been disputed by Lander and Kruglyak (1995), Visscher and Haley (2001) and Siegmund (2001), although none of these papers contains a systematic analysis.

The claim of Terwilliger *et al.* (1997) that wider peaks are more likely to contain the gene of interest than narrower peaks of similar height is based on the length-biased sampling principle in renewal theory. Their argument has been refuted by Visscher and Haley (2001), who notice that the length differences observed by Terwilliger *et al.* are due to the fact that they are comparing the width of the peak at a fixed point with all peaks of similar height and that the same differences would be observed for any other prespecified locus regardless of whether it is linked to the trait or not. Both papers illustrate their respective reasoning with simulations, but neither suggests a test for linkage and evaluates its power.

In this paper we formally define a statistical test based on moving averages for IBD proportions of affected pairs of relatives and derive approximations for its significance level and power. We use these results to show that in the situation envisaged by Hoh and Ott (2000), the scan statistics do not have increased power. We also study an alternative situation, where two trait loci are closely linked and one would expect to find a "wide peak."

In human genetics the possibility of more than one trait-locus on a chromosome has received relatively little attention. An exception is Farrall (1997). However, some multigene families, presumably derived from one original gene via various mutations during the course of evolution, suggest that genes affecting the same trait may be located near each other (cf. Strachan and Read, 1996, p. 187 ff.). Our results suggest that the use of scan statistics may help to detect linkage in these cases, although the increase in power is modest.

In order to study what we think is the essence of the problem, we assume data from a dense set of completely informative markers and samples that are large enough that normal approximations are valid. We discuss later how the results change for a discrete set of markers. For markers that are only partially informative, the overall performance of each statistic will be degraded, but their relative merits should not change appreciably. To simplify our exposition of the genetic background, we analyze primarily the simple case of independent half-sibling pairs and discuss later the minor changes that are required for the more common case of affected sib pairs and for other types of relatives.

Our analysis in this paper is confined to a simple genome scan designed to detect relatively weak linkage signals. A complementary conditional approach would be useful in situations where there is an easily detected trait-locus with a large effect and another tightly linked trait-locus with a minor effect, which may be masked by the major gene. We will discuss this topic in a future paper.

2. MODELS

For a pair of half-siblings and a locus situated on chromosome c at position t we define the variable $D_{c,t}$ as

$$D_{c,t} = \begin{cases} 1, & \text{if the half-sibling pair is IBD at locus } t, \\ 0, & \text{otherwise.} \end{cases}$$

The location t indicates genetic distance in centimorgans (cM) from a fixed end of the chromosome. Let L_c denote the genetic length of chromosome c. For a randomly chosen pair of half-siblings, Mendel's laws imply that

$$P[D_{c,t} = 1] = P[D_{c,t} = 0] = \frac{1}{2} \tag{2.1}$$

for each c and t in $[0, L_c]$, and that for chromosomes $c \neq c'$, the variables $D_{c,t}$ and $D_{c',s}$ are independent for all t in $[0, L_c]$ and s in $[0, L_{c'}]$. If we use the Haldane mapping function, which specifies that the

number of crossovers during meiosis follows a Poisson process, then for loci located at position t and s in the same chromosome,

$$P[D_{c,s} = 1|D_{c,t} = 1] = P[D_{c,s} = 0|D_{c,t} = 0] = \frac{1 + e^{-\beta|t-s|}}{2}. \tag{2.2}$$

The value of the parameter β is determined by the kinship of the pair of relatives. For the case of half-siblings, whose IBD status is determined by two meioses, the value is $\beta = 4$ if distances are measured in Morgans and $\beta = 0.04$ if the units of genetic distance are cMs.

Let us suppose now that our pair is chosen among the population of pairs of half-siblings that are both affected by a particular trait of interest. We will consider three possible situations:

(A) If there is no locus in the genome predisposing for the trait, or if all trait loci have very weak effects that cannot be detected without an unacceptable level of false-positive errors, then (2.1) will be approximately valid for all loci and chromosomes.

(B) If at chromosome c_0 there is a locus τ which predisposes for inheritance of the trait and there is no other trait-locus on the same chromosome, then

$$P[D_{c_0,\tau} = 1] = \frac{1 + \alpha}{2} > \frac{1}{2}. \tag{2.3}$$

The parameter $\alpha > 0$ measures the increase of likelihood of sharing an allele IBD at the trait-locus for pairs of half-siblings who share the trait of interest. A description of α in terms of allele frequencies and penetrances of the trait has been given by a number of authors, e.g. Risch (1990a,b), Feingold *et al.* (1993), Dupuis *et al.* (1995). Under the Haldane mapping function, for a locus at position t on the same chromosome as τ,

$$P[D_{c_0,t} = 1] = \frac{1 + \alpha\, e^{-\beta|\tau-t|}}{2},$$

while (2.1) continues to hold for loci located in chromosomes that do not contain any trait loci.

(C) If chromosome c_0 contains two trait loci that do not interact, located at positions τ_1 and τ_2, then the probability of identity by descent at a locus $t \in c_0$ can be expressed as

$$P[D_{c_0,t} = 1] = \frac{1 + \alpha_1\, e^{-\beta|\tau_1-t|} + \alpha_2\, e^{-\beta|\tau_2-t|}}{2}, \tag{2.4}$$

with $\alpha_1, \alpha_2 > 0$ (see Dupuis *et al.*, 1995) and (2.1) is again valid for t in chromosomes without trait loci.

The overall goal is to determine whether the trait is of a genetic nature or not and to establish the number of genes affecting the trait and their approximate genomic locations.

A similar approach can be applied to other kinds of pairs of relatives. Depending on the type of kinship, different modifications are required. The most important case is pairs of siblings, for which again $\beta = 0.04$ (see, for example Risch 1990a,b) and the essential ingredient of a commonly used statistic is $M_{c,t}$, which denotes the number of alleles, 0, 1 or 2, shared IBD by a sib pair on chromosome c at locus t. Note that for each sib pair this statistic can be expressed as the sum of two terms of the form of $D_{c,t}$, if we think of each pair of siblings as two pairs of half-siblings, one related through their maternally and the other through their paternally inherited chromosomes. On unlinked chromosomes these half-sib pairs are independent; on linked chromosomes they behave conditionally independently given the IBD counts at trait loci. As a consequence, the asymptotic results given below for the significance level and power do

not change, although the representation of the overall non-central parameter of the test statistic in terms of a genetically interpretable parameter, denoted by α in (2.3), does change.

3. METHODS

Here we discuss the single locus search approach, i.e. we restrict our attention to methods based on the model with a unique trait-locus described in (B) of Section 2. Such methods can also be useful to detect the presence of several trait loci (cf. Section 6). We assume that we observe identity by descent data from N independent pairs of affected half-siblings and introduce the new parameter $\xi = \sqrt{N}\alpha$. For the model defined in (B), to test for genetic linkage means to test the null hypothesis of non-existence of a trait-locus, $H_0\colon \xi = 0$, versus the alternative of genetic linkage, $H_1\colon \xi > 0$, at some trait-locus τ. The score statistic (see Cox and Hinkley, 1974, pp. 313–331) for testing the hypothesis of no trait-locus at a putative trait-locus t on chromosome c is

$$V_{c,t} = N^{-1/2} \sum_{j=1}^{N} \left(2\, D_{c,t}^{j} - 1\right), \tag{3.1}$$

where $D_{c,t}^{j}$ denotes the IBD indicator at locus t on chromosome c for the pair j; the hypothesis of no trait-locus at such a position is rejected when $V_{c,t}$ is large. Since the position of the possible trait-locus is unknown, the test for linkage is based on the maximum of $V_{c,t}$ over the entire genome, and H_0 is rejected if we observe large enough values of

$$\max_{\text{genome}} V_{c,t} = \max_{1 \leqslant c \leqslant C} \; \max_{0 \leqslant t \leqslant L_c} V_{c,t},$$

where C denotes the number of pairs of chromosomes.

Alternatively, we can follow the proposal of Hoh and Ott (2000) and use the scan statistic based on $V_{c,t}$ to test whether there exists a trait-locus or not. For dense markers the scan of bandwidth $\epsilon > 0$ is defined as

$$S_{c,t}^{\epsilon} = \frac{1}{2\epsilon} \int_{t-\epsilon}^{t+\epsilon} V_{c,s}\, \mathrm{d}s. \tag{3.2}$$

In Appendix A.1 of the supporting material (see *Biostatistics* online) we show that after standardizing to have unit variance under the null hypothesis, the scan statistic becomes

$$Q_{c,t}^{\epsilon} = \kappa^{\epsilon} \beta \int_{t-\epsilon}^{t+\epsilon} V_{c,s}\, \mathrm{d}s, \tag{3.3}$$

where $\kappa^{\epsilon} = [2(2\beta\epsilon - 1 + \mathrm{e}^{-2\beta\epsilon})]^{-1/2}$. This statistic is a smoothed version of $V_{c,t}$ which combines the information coming from several contiguous markers. Note that, for a chromosome of genetic length L_c, $Q_{c,t}^{\epsilon}$ is only defined at positions $t \in [\epsilon, L_c - \epsilon]$. A test based on this scan statistic should reject the hypothesis of no linkage when

$$\max_{\text{genome}} Q_{c,t}^{\epsilon} = \max_{1 \leqslant c \leqslant C} \; \max_{\epsilon \leqslant t \leqslant L_c - \epsilon} Q_{c,t}^{\epsilon}$$

is sufficiently large.

Our goal is to compare the power of the tests based on $Q_{c,t}^{\epsilon}$ and on $V_{c,t}$ to detect linkage under situations (B) and (C).

4. Significance levels and power

4.1 *Significance levels*

The first step is to determine which values of the maximum of each statistic are large enough to imply evidence of genetic linkage. To this end, we evaluate the genome-wide false-positive error rates

$$P_0\left[\max_{\text{genome}} V_{c,t} > b\right] \quad \text{and} \quad P_0\left[\max_{\text{genome}} Q^{\epsilon}_{c,t} > b\right],$$

where the subscript 0 indicates that the probabilities are computed under $\xi = 0$. Because of the independent assortment of the chromosomes during meiosis,

$$P_0\left[\max_{\text{genome}} V_{c,t} > b\right] = 1 - \prod_{c=1}^{C} P_0\left[\max_{0\leqslant t\leqslant L_c} V_{c,t} \leqslant b\right],$$

and similarly

$$P_0\left[\max_{\text{genome}} Q^{\epsilon}_{c,t} > b\right] = 1 - \prod_{c=1}^{C} P_0\left[\max_{\epsilon\leqslant t\leqslant L_c-\epsilon} Q^{\epsilon}_{c,t} \leqslant b\right];$$

hence, it is sufficient to look at the false-positive error rate of individual chromosomes. In what follows, we will omit the index c in $V_{c,t}$, $S^{\epsilon}_{c,t}$, $Q^{\epsilon}_{c,t}$ and L_c and write V_t, S^{ϵ}_t, Q^{ϵ}_t and L, respectively, when we consider only one chromosome.

Let us assume that data are available from a dense set of fully informative markers. Under this assumption, Feingold *et al.* (1993) suggest the following approximation to the probability that the maximum of V_t over a chromosome of length L exceeds a threshold b when there is no trait-locus:

$$P_0\left[\max_{0\leqslant t\leqslant L} V_t > b\right] \simeq 1 - \Phi(b) + \beta L b \phi(b), \tag{4.1}$$

where ϕ and Φ denote the standard normal density and distribution function, respectively.

We show in Appendix A.2 (see online supporting material) that as a consequence of Rice's formula for the expected number of upcrossings of a level by a smooth random process (cf. Leadbetter *et al.*, 1983 or Davies, 1987), the false-positive rate of the test based on the smoothed statistic Q^{ϵ}_t is approximately given by

$$P_0\left[\max_{\epsilon\leqslant t\leqslant L-\epsilon} Q^{\epsilon}_t > b\right] \simeq 1 - \Phi(b) + \kappa^{\epsilon}\beta(L - 2\epsilon)\phi(b)\sqrt{\frac{1 - \mathrm{e}^{-2\beta\epsilon}}{\pi}}. \tag{4.2}$$

Formulae (4.1) and (4.2) are Gaussian approximations based on the central limit theorem, and hence they are valid only for large sample sizes.

As a numerical illustration, we consider a genome consisting of 23 pairs of chromosomes of average length 140 cM. In order to obtain the conventional genome-wide significance level of 0.05 we need a significance level of about 0.0022 for each chromosome. For the non-smoothed statistic V_t according to formula (4.1) this corresponds to the threshold $b = 4.08$. By using formula (4.2) we find that the thresholds corresponding to the standardized smoothed statistic Q^{ϵ}_t with $\epsilon = 5$, 10, 15, 20, 25, 30, 35 and 40 are $b = 3.68$, 3.56, 3.48, 3.14, 3.36, 3.30, 3.25 and 3.20, respectively. Note that as ϵ increases the appropriate threshold to provide the same false-positive error rate becomes smaller. This is in part a consequence of the range of values of t over which we take maxima becoming narrower but also of the larger degree of smoothness.

4.2 *Power when there is one trait-locus*

Suppose that the model with a unique trait-locus described in case (B) is valid for some locus τ and some $\xi > 0$. The approximation to the power of the test based on the score statistic V_t suggested by Feingold *et al.* (1993) for large b, ξ and N is

$$P_\xi\left[\max_{0\leqslant t\leqslant L} V_t > b\right] \simeq 1 - \Phi(b-\xi) + \phi(b-\xi)\left(\frac{2}{\xi} - \frac{1}{b+\xi}\right). \tag{4.3}$$

In Appendix A.3 of the online supporting material we obtain the following asymptotic approximation to the power to detect linkage of the test based on the scan statistic

$$P_\xi\left[\max_{\epsilon\leqslant t\leqslant L-\epsilon} Q_t^\epsilon > b\right] \simeq 1 - \Phi(b - m_\xi^\epsilon) + \frac{\phi(b-m_\xi^\epsilon)}{b-m_\xi^\epsilon}\left(\sqrt{1 + \frac{m_\xi^\epsilon(b-m_\xi^\epsilon)\left(1+\mathrm{e}^{\beta\epsilon}\right)}{2\xi^2}} - 1\right), \tag{4.4}$$

where $m_\xi^\epsilon = 2\xi\kappa^\epsilon\left(1-\mathrm{e}^{-\beta\epsilon}\right)$ is the expected value of the scan statistic at the trait-locus. The first term on the right-hand sides of (4.3) and (4.4) is the probability that the statistic exceeds b at the trait-locus, while the second term approximates the probability of being below the threshold b at $t = \tau$ but above b at some other locus t close to τ. Simulations show that (4.4) is a good approximation for large sample sizes, although the second term should be divided by 2 if τ is close to either end of the chromosome.

To obtain some insight into how the power changes as the bandwidth ϵ increases, we have numerically evaluated the case $\xi = N^{1/2}\alpha = 5$ for the genome described above and for an overall significance level of 0.05. Table 1 displays the approximated values of the power of the tests based on Q_t^ϵ for $\epsilon = 0, 5, 10, 15, 20, 25, 30, 35$ and 40. Note that the case $\epsilon = 0$ corresponds to the non-smoothed statistic V_t and its approximated power has been calculated using (4.3), while (4.4) has been used for the other cases. The table shows that the power to detect linkage slowly decreases as ϵ increases, but the loss in power when using Q_t^ϵ with ϵ in the range $(0, 25]$ instead of V_t is not large. Similar results were obtained for other values of ξ.

Table 1. *Approximate power for the case of one trait-locus with effect* $\xi = 5$*

Bandwidth (ϵ)	Statistic	Threshold (b)	Power
0	V_t	4.08	0.90
5	Q_t^5	3.68	0.90
10	Q_t^{10}	3.56	0.89
15	Q_t^{15}	3.48	0.88
20	Q_t^{20}	3.41	0.86
25	Q_t^{25}	3.36	0.85
30	Q_t^{30}	3.30	0.82
35	Q_t^{35}	3.25	0.81
40	Q_t^{40}	3.20	0.79

*For a genome of 23 pairs of chromosomes of average length 140 cM and a genome-wide significance level of 0.05.

Table 2. *Approximate power for the case of two trait loci with effects* $\xi_1 = 3$ *and* $\xi_2 = 2$*

ϵ	b	$\lvert\tau_2 - \tau_1\rvert = 10$	$\lvert\tau_2 - \tau_1\rvert = 20$	$\lvert\tau_2 - \tau_1\rvert = 30$	$\lvert\tau_2 - \tau_1\rvert = 40$
0	4.08	0.824	0.727	0.600	0.495
5	3.68	0.840	*0.730*	*0.624*	*0.531*
10	3.56	0.854	0.752	*0.646*	*0.543*
15	3.48	0.856	0.766	0.660	*0.557*
20	3.41	0.848	0.789	0.667	0.581
25	3.36	0.834	0.789	0.705	0.563
30	3.30	0.819	0.786	0.724	0.624
35	3.25	0.804	0.777	0.729	0.653
40	3.20	0.790	0.767	0.728	0.669

*For a genome of 23 pairs of chromosomes of average length 140 cM and an overall significance level of 0.05.
Numbers in italics correspond to situations in which $0 < \epsilon < \epsilon^*$. These values have been computed by Monte Carlo simulations since the approximations do not apply.

4.3 *Power when there are two linked trait loci*

We now assume that chromosome c_0 contains two trait loci at unknown positions τ_1 and τ_2 with additive effects α_1 and α_2, as in the model described in case (C). We set $\xi_i = N^{1/2}\alpha_i$ for $i = 1, 2$. In Appendix A.4.1 (see online supporting material) we give an asymptotic approximation for the power of the test based on the score statistic, $P_{\xi_1,\xi_2,\tau_1,\tau_2}[\max_{0\leqslant t\leqslant L} V_t > b]$. The approximation has a rather complicated expression and requires numerical computation of several integrals (see Proposition 3 in the online supporting material).

In Appendix A.4.2 (see online supporting material) we obtain an approximate formula for the power of the scan statistic, $P_{\xi_1,\xi_2,\tau_1,\tau_2}[\max_{\epsilon\leqslant t\leqslant L-\epsilon} Q_t^\epsilon > b]$, whose expression varies depending on the relation between all the parameters (see Proposition 5 in the online supporting material). This approximation is quite accurate provided that $\epsilon \geqslant \epsilon^*$, where ϵ^* is half the genetic distance between the two trait loci plus a term which depends on the ratio ξ_1/ξ_2 (see A.25 in Appendix A.4.2 of the online supporting material).

Table 2 displays the approximated power to detect linkage corresponding to the case $\xi_1 = 3$ and $\xi_2 = 2$ for different distances between τ_1 and τ_2 and the same bandwidths ϵ as in Table 1. We have used the result in Proposition 3 (see online supporting material) to evaluate the power corresponding to $\epsilon = 0$ and the formulae in Proposition 5 for the cases with $\epsilon \geqslant \epsilon^*$. The numbers in italics correspond to situations in which $0 < \epsilon < \epsilon^*$. Since our approximations do not apply in such cases, these values have been computed by Monte Carlo simulations.

The table shows that in cases with two trait loci on a chromosome the smoothed statistic Q_t^ϵ provides a modest increase in power. The bandwidth ϵ that gives the largest power depends on the distance between the two trait loci. In all cases, the maximum power corresponds to the test based on the scan statistic with a bandwidth close to the distance between the two trait loci.

5. Corrections for discrete sets of markers

In previous sections we have assumed that a completely dense set of markers is available, namely that it is possible to look at the IBD status of a pair of relatives at every location along the genome. In practice, the information about the IBD status is limited to a discrete set of genetic markers.

Since the sample paths of the scan statistics are smooth, the differences between a dense and a discrete set of markers are much smaller for Q_t^ϵ than for the non-smoothed score statistic, V_t. Hence, for the tests

based on the scan statistics Q_t^{ξ} it is usually not necessary to apply corrections for discrete markers unless the markers are very widely spaced. For example for an intermarker distance of 5 cM and $\epsilon = 20$, in which case the scan statistic involves moving averages of 8 markers, simulations indicate that the threshold and power when $\xi = 5$ are the same as the values given in Table 1, and there is almost no difference in power whether the trait-locus is at a marker or midway between markers. For an intermarker distance of 10 cM and $\epsilon = 25$, in which case the scan statistic involves moving averages of 5 markers, simulations indicate a threshold of 3.34 and power of 0.85 or 0.82 for the trait-locus being at or midway between two markers, respectively. Thus, it appears that scan statistics with a moderate to large window size have sufficiently smooth behavior that corrections for discrete sampling are rarely required.

Since the sample paths of V_t fluctuate much more rapidly, the formulae for the false-positive error rate and the power of the test based on the non-smoothed score statistic V_t need to be modified. Feingold *et al.* (1993) give the following approximation for the false-positive error rate for equally spaced markers at intermarker distance Δ

$$P_0\left[\max_{0\leqslant i\Delta\leqslant L} V_{i\Delta} > b\right] \simeq 1 - \Phi(b) + \beta L b\, \phi(b)\, \nu[b(2\beta\Delta)^{1/2}], \tag{5.1}$$

where $\nu(x)$ is the special function defined in Siegmund (1985) and in the range $0 < x < 2$ is very well approximated by $\exp(-\varrho x)$ with $\varrho \simeq 0.583$. For the genome of our example and intermarker distances $\Delta = 0.1, 1, 5$ and 10, the 0.05 false-positive approximate thresholds are $b = 4.03$, 3.91, 3.73 and 3.60, respectively.

Siegmund (1998) derives the following approximate formula for the power of the test based on V_t for cases with one trait-locus provided that the trait-locus τ is itself a marker locus,

$$P_{\xi}\left[\max_{0\leqslant i\Delta\leqslant L} V_{i\Delta} > b\right] \simeq 1 - \Phi(b-\xi) + \phi(b-\xi)\left(\frac{2\nu}{\xi} - \frac{\nu^2}{b+\xi}\right), \tag{5.2}$$

where $\nu = \nu[b(2\beta\Delta)^{1/2}]$ as above. For our example with $\xi = 5$ and a genome-wide significance level of 0.05, if τ is one of the genetic markers, the power corresponding to equally spaced markers at intermarker distances $\Delta = 0.1, 1, 5$ and 10 has approximate values 0.90, 0.90, 0.92 and 0.93, respectively. A somewhat more complicated formula applies when the trait-locus is between markers. When it is exactly midway between markers, the corresponding values of the power are approximately 0.89, 0.89, 0.87 and 0.82, respectively. We see then that the power of the unsmoothed statistic is more sensitive to the location of the trait-locus with respect to its flanking markers.

Table 3. *Approximated power of the test based on the non-smoothed statistic V_t for discrete equally spaced markers at intermarker distance Δ when there are two trait loci with effects $\xi_1 = 3$ and $\xi_2 = 2$**

Δ	b	$\lvert\tau_2 - \tau_1\rvert = 10$	$\lvert\tau_2 - \tau_1\rvert = 20$	$\lvert\tau_2 - \tau_1\rvert = 30$	$\lvert\tau_2 - \tau_1\rvert = 40$
0	4.08	0.824	0.727	0.600	0.497
0.1	4.03	0.843	0.728	0.612	0.497
1	3.91	0.850	0.737	0.613	0.509
5	3.73	0.869	0.759	0.639	0.538
10	3.60	0.890	0.787	0.672	0.574

*For a genome of 23 pairs of chromosomes of average length 140 cM and an overall significance level of 0.05.

In Proposition 4 (see online supporting material) we give a modified version of the approximation in Proposition 3 for the power when the chromosome contains two trait loci, which is valid for discrete equally spaced markers. We assume that both trait loci are located at the sites of markers. Table 3 shows the approximate power for discrete markers for $\Delta = 0.1, 1, 5$ and 10 corresponding to our example with $\xi_1 = 3$ and $\xi_2 = 2$ and a false-positive error rate of 0.05, provided that τ_1 and τ_2 are both marker loci.

We see from Table 3 that the power increases slightly with Δ, but as in the case of a single trait-locus, we expect it to decrease by roughly an equal amount when the trait loci are between genetic markers.

6. Discussion

In this paper we have examined some aspects of scan statistics (moving averages of the values at contiguous markers of an initial statistic). Assuming that data about the identity by descent status of pairs of affected relatives are available, we have compared the performance of the test based on the classical score statistic with that of the scan statistic.

We have determined the thresholds to control the genome-wide false-positive rate of the tests based on scan statistics and have derived approximate formulae for the power to detect linkage in situations with one trait-locus and with two trait loci on a chromosome. We have also presented approximations to the power of the test based on the non-smoothed score statistic when there are two linked genes affecting the trait for both dense and discrete markers.

A numerical evaluation of these formulae indicates that in the case of a single trait-locus, the smoothed statistic has slightly less power than the original statistic. This contradicts the suggestion (Hoh and Ott, 2000) that power would increase because of the argument that "true peaks are wider than false peaks" advanced by Terwilliger *et al.* (1997). In the case that there are two trait loci in the same chromosomal region, the smoothed statistic provides a modest increase in the power to detect linkage. Since computation of the scan statistic requires very little extra effort, we can say that in order to increase the chances to detect linkage it may be worthwhile to consider the tests based on the scan statistic. Moreover, a comparison of the p-values from the test based on the non-smoothed score statistic with those of the tests based on the scan statistic with a few bandwidths may be useful to discriminate between cases with only one trait-locus versus cases with more than one trait-locus on a given chromosome.

Although the moving average statistic appears to lose a small amount of power in comparison with the score statistic when there is only one linked gene, its power appears to be less sensitive to the location of the gene with respect to flanking markers.

In cases with two linked trait loci, the bandwidth that gives the largest power depends on the distance between the loci. The numerical results indicate that the maximum power corresponds to the test based on the scan statistic with a bandwidth close to the distance between the two trait loci. This suggests (as do Hoh and Ott, 2000) that the bandwidth ϵ should be selected adaptively. Siegmund and Worsley (1995) describe appropriate modifications for the p-value. Since a higher threshold is required, and since even an optimal choice of ϵ produces only a moderate increase in power, it is not clear how useful this idea is, although it might also allow one to get some idea of the distance between linked trait loci.

In this paper we have considered smoothing with the uniform kernel, so our results could be directly compared with those of Hoh and Ott (2000). It is also possible, and perhaps advantageous, to consider smoothing with other kernels, e.g. a Gaussian kernel, which would produce smoother sample paths and not lead to discontinuous behavior in the approximation to the power function that occurs with the uniform kernel when ϵ is half the distance between the two trait loci.

We have assumed throughout that markers are completely informative. When markers are less than fully informative, multi-point analysis to maximize information recovery is itself a kind of smoothing, but it is non-linear smoothing conducted at the level of the pedigree, not at the level of the statistic. In this

case, both statistics will lose some power, although we expect their comparative value to remain roughly the same. If one uses only single point analysis, i.e. IBD status at any particular marker is inferred only from genotypes and allele frequencies for that marker, then the linear smoothing discussed in this paper is a weak form of multi-point analysis, which will improve power to detect genes located near markers with low information content.

The method described here seems to be a useful tool in cases where linked trait loci have a modest effect and none of them can easily be detected by the standard approach. A complementary situation happens when one of the two linked trait loci has a strong effect and is easy to detect, while the effect of the second one is relatively small and is masked by the major gene. In a future paper we will analyze how a conditional approach for sequential detection of the trait loci performs in such situations.

ACKNOWLEDGMENTS

This work was started when S. Hernández was visiting Stanford University supported by a EURANDOM fellowship. The research of D. O. Siegmund was partially supported by NIH Grant HG 00-848. The authors are grateful to Dorret Boomsma and Chris Klaassen for helpful discussions.

REFERENCES

COX, D. R. AND HINKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

DAVIES, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74**, 33–43.

DUPUIS, J., BROWN, P. AND SIEGMUND, D. (1995). Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent. *Genetics* **140**, 843–856.

FARRALL, M. (1997). Affected sibpair linkage tests for multiple linked susceptibility genes. *Genetic Epidemiology* **14**, 103–115.

FEINGOLD, E., BROWN, P. AND SIEGMUND, D. (1993). Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. *American Journal of Human Genetics* **53**, 234–251.

HOH, J. AND OTT, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proceedings of the National Academy of Sciences* **95**, 9615–9617.

LANDER, E. AND KRUGLYAK, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241–247.

LEADBETTER, M. R., LINDGREN, G. AND ROOTZN, H. (1983) *Extremes and Related Properties of Random Sequences and Processes*. New York: Springer-Verlag.

RISCH, N. (1990a). Linkage strategies for genetically complex traits I. Multilocus models. *American Journal of Human Genetics* **46**, 222–228.

RISCH, N. (1990b). Linkage strategies for genetically complex traits II. The power of affected relative pairs. *American Journal of Human Genetics* **46**, 229–241.

SIEGMUND, D. (1985). *Sequential Analysis*. New York: Springer.

SIEGMUND, D. (1998). Genetic linkage analysis: an irregular statistical problem. *Documenta Mathematica*. Extra Volume ICM III, 257–266.

SIEGMUND, D. (2001). Is peak height sufficient? *Genetic Epidemiology* **20**, 403–408.

SIEGMUND, D. AND WORSLEY, K. J. (1995). Testing for a signal with unknown location and scale in a stationary random field. *The Annals of Statistics* **23**, 608–639.

STRACHAN, T. AND READ, A. P. (1996). *Human Molecular Genetics*. Oxford: BIOS Scientific Publishers.

TERWILLIGER, J. D., SHANNON, W. D., LATHROP, G. M., NOLAN, J. P., GOLDIN, L. R., CHASE, G. A. AND WEEKS, D. E. (1997). True and false positive peaks in genomewide scans: applications of length-biased sampling to linkage mapping. *American Journal of Human Genetics* **61**, 430–438.

VISSCHER, P. AND HALEY, C. (2001). True and false positive peaks in genomewide scans: the long and the short of it. *Genetic Epidemiology* **20**, 409–414.

[*Received February 12, 2004; first revision July 19, 2004; second revision September 27, 2004; accepted for publication November 11, 2004*]