On the Power for Linkage Detection Using Tests Based on Scan Statistics

Sonia Hernández* and David O. Siegmund †

Abstract

We analyze some aspects of Scan statistics, which have been proposed to help detect weak signals in linkage analysis. For dense markers we derive approximations for the thresholds to control the genome wide false positive rate and for the power of a test based on moving averages of the identity-by-descent (IBD) allele sharing proportions for pairs of relatives at several contiguous markers. We use these results, which we confirm by simulation, to show that when there is a single trait gene on a chromosome this scan statistic is generally slightly less powerful than the customary allele sharing statistic, but if two genes having a moderate affect on the same trait lie close to each other on the same chromosome, this test can be more powerful than that based on the original statistic.

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 odds (LOD) score 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 systematic study of these statistics, in an application to autism families they find a region that was missed with the standard approach. The value of this approach has been disputed by Lander and Kruglyak (1995) and Siegmund (2001), although neither of these papers contains a sytematic analysis.

The purpose of this paper is twofold: (i) for dense markers we derive approximations for the significance level and power of tests based on scan statistics for IBD proportions of affected relative pairs and use these results to show that in the situation envisaged by Hoh and Ott (2001) the scan statistics do not have increased power, and (ii) we study an alternative situation, where two trait loci are closely linked, where one would expect to find a "wide peak."

In human genetics the possibility of more than one trait locus on the same chromosome has received relatively little attention. An exception can be found in 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

*Eurandom, P.O. Box 513 - 5600 MB Eindhoven, The Netherlands. (hernandez@eurandom.tue.nl)
†Department of Statistics, Stanford University

be located near each other (cf. Strachan and Read (1996)). Our results suggest that using the scan statistic may help to detect linkage in these cases, although the increase in power is modest.

In order to get to 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 and for markers that are only partially informative. To simplify our exposition of the genetics background, we discuss primarily the simple case of independent half-sibling pairs, and discuss later how the minor changes that occur 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

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

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

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

for all 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,

$$P[D_{c,s} = 1 | D_{c,t} = 1] = P[D_{c,s} = 0 | D_{c,t} = 0] = \frac{1 + e^{-\beta|t-s|}}{2}, \tag{3}$$

where $\beta = 0.04$ for half sibling pairs if the units of genetic distance are centimorgans.

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 cannot be detected without an unacceptable level of false positive errors, then (2) 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

$$P[D_{c_0,\tau} = 1] = \frac{1 + \alpha}{2} > \frac{1}{2}. \tag{4}$$

The parameter $\alpha > 0$ measures the increase of likelihood of sharing an allele identical by descent 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 give by a number of authors, e.g., Risch (1990a, 1990b), 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) 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

$$P[D_{c_0,t} = 1] = \frac{1 + \alpha_1\, e^{-\beta|\tau_1 - t|} + \alpha_2\, e^{-\beta|\tau_2 - t|}}{2}, \tag{5}$$

with $\alpha_1, \alpha_2 > 0$ (see Dupuis et al., 1995) and (2) is again valid for t in chromosomes without trait-loci.

The main goal is to determine whether the trait is of a genetic nature or not, that is to discriminate between cases (B) or (C) versus case (A). If a genetic nature is established for the trait it would also be of interest to clarify whether there is only one trait-locus in chromosome c_0 –case (B)– or whether there are several linked trait-loci –case (C)– and to find the approximate location of the trait-locus or loci.

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 $\beta = 0.04$ and the essential ingredient of a commonly used statistic is $M_{c,t}$, which denotes the number of allels, 0,1,or 2, shared IBD by a sib pair on the cth chromosome at the locus t. Note that 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 alleles. On unlinked chromosomes these half sib pairs are independent, and on linked chromsomes 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 over all noncentrality parameter of the test statistic in terms of a genetically interpretable parameter—which we denoted α in (4)—does change. See, for example, Risch (1990a,b).

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 the previous section. 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, $H_0 : \xi = 0$, versus the alternative of genetic linkage, $H_1 : \xi > 0$, at some trait-locus τ. The score statistic (see for instance Cox and Hinkley, 1974) for testing the hypothesis of no trait-locus at a putative trait locus t on chromosome c is

$$V_{c,t} = N^{-1/2} \sum_{j=1}^{N} \Big(2\, D^j_{c,t} - 1\Big), \tag{6}$$

where $D^j_{c,t}$ 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

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

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

$$S^{\epsilon}_{c,t} = \frac{1}{2\epsilon} \int_{t-\epsilon}^{t+\epsilon} V_{c,s}\, ds, \tag{7}$$

which after standardizing to have unit variance under the null hypothesis becomes (see Appendix I)

$$Q^{\epsilon}_{c,t} = \kappa^{\epsilon}\beta \int_{t-\epsilon}^{t+\epsilon} V_{c,s}\, ds, \tag{8}$$

where $\kappa^{\epsilon} = [2(2\beta\epsilon - 1 + e^{-2\beta\epsilon})]^{-1/2}$. 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, $Q^{\epsilon}_{c,t}$ 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

$$\max_{\text{genome}} Q^{\epsilon}_{c,t} = \max_{1\le c\le C} \max_{\epsilon\le t\le L_c-\epsilon} Q^{\epsilon}_{c,t}$$

is sufficiently large.

We want to compare the power of the tests based on $Q^{\epsilon}_{c,t}$ and on $V_{c,t}$ to detect linkage under situations (B) and (C).

4. Significance levels and power. The first step is to determine which values of the maximum of each statistic are big enough to state evidence for genetic linkage, which implies that the genome-wide false-positive error rates

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

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

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

and similarly

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

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 $V_t, S^{\epsilon}_t, Q^{\epsilon}_t$, 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

$$P_0\left[\max_{0\le t\le L} V_t > b\right] \simeq 1 - \Phi(b) + \beta L b \phi(b), \tag{9}$$

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

We show in Appendix II that as a consequence of Rice's formula for the expected number of upcrossings of a level by a smooth random process the false-positive rate of the test based on the smoothed statistic Q_t^ϵ is approximately given by

$$P_0\left[\max_{\epsilon\le t\le L-\epsilon} Q_t^\epsilon > b\right] \simeq 1-\Phi(b)+\kappa^\epsilon\beta(L-2\epsilon)\phi(b)\sqrt{\frac{1-e^{-2\beta\epsilon}}{\pi}}. \tag{10}$$

Formulae (9) and (10) 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 (9) this corresponds to the threshold $b = 4.08$. By using formula (10) we find that the thresholds corresponding to the smoothed statistic Q_t^ϵ 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.

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 position 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

$$P_\xi[\max_{0\le t\le L} V_t > b] \simeq 1-\Phi(b-\xi)+\phi(b-\xi)\left(\frac{2}{\xi}-\frac{1}{b+\xi}\right). \tag{11}$$

In Appendix III we obtain the following asymptotic approximation to the power to detect linkage of the test based on the scan statistic

$$\begin{aligned} &P_\xi\left[\max_{\epsilon\le t\le L-\epsilon} Q_t^\epsilon > b\right] \\ &\simeq\ 1-\Phi(b-m_\xi^\epsilon)+\frac{\phi(b-m_\xi^\epsilon)}{b-m_\xi^\epsilon}\left(\sqrt{1+\frac{m_\xi^\epsilon(b-m_\xi^\epsilon)\,(1+e^{\beta\epsilon})}{2\xi^2}}-1\right), \end{aligned} \tag{12}$$

where $m_\xi^\epsilon = 2\xi\kappa^\epsilon\left(1-e^{-\beta\epsilon}\right)$ is the expected value of the scan statistic at the trait-locus. The first term on the right-hand side of (11) and of (12) equal the probability that the scan statistic exceeds b at the trait-locus, while the last 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 this approximation is quite good 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 (11), while (12) 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 $\epsilon \in (0, 25]$ instead of V_t is not large. Similar results were obtained for other values of ξ.

Table 1. Approximations to the power for 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

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 the first subsection of Appendix IV we give an asymptotic approximation for the power of the test based on the score statistic, $P_{\xi_1,\xi_2,\tau_1,\tau_2}[\max_{0\le t\le L} V_t > b]$. The approximation has a rather complicated expression and requires numerical computation of several integrals (see Proposition 3).

In the second subsection of Appendix IV we obtain an explicit approximation for the power of the scan statistic, $P_{\xi_1,\xi_2,\tau_1,\tau_2}[\max_{\epsilon\le t\le L-\epsilon} Q_t^\epsilon > b]$, whose expression varies depending on the relation between all the parameters (see Proposition 4). This approximation is quite accurate provided that $\epsilon \ge \epsilon^*$, 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 (35) in Appendix IV.

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 to evaluate the power corresponding to $\epsilon = 0$ and the formulae in Proposition 4 for the cases with $\epsilon \ge \epsilon^*$. 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.

Table 2. Approximations to the power for 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

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 set of markers and 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^ϵ 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 $\epsilon = 20$, so 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 $\epsilon = 25$, so the scan statistic involves moving averages of 5 markers, simulations indicate a threshold of 3.34, and power of 0.85 or 0.82 according as the trait locus is at or midway between two markers. 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

$$P_0\left[\max_{0\le i\Delta\le L} V_{i\Delta} > b\right] \simeq 1 - \Phi(b) + \beta L b\,\phi(b)\,\nu[b(2\beta\Delta)^{1/2}], \tag{13}$$

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

$$P_\xi\left[\max_{0\le i\Delta\le L} V_{i\Delta} > b\right] \simeq 1 - \Phi(b-\xi) + \phi(b-\xi)\left(\frac{2\nu}{\xi} - \frac{\nu^2}{b+\xi}\right), \tag{14}$$

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, then 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.

In Proposition 3* we give a modified version valid for discrete equally spaced markers of the approximation in Proposition 3 for the power when the chromosome contains two trait loci. 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.

Table 3. Approximations to the power of the test based on 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	0.843	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

We see in the examples 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 approximated formulae for the power to detect linkage in situations with one trait-locus and with two trait-loci in a chromosome. We have also presented approximations to the power of the test based on the non-smoothed score statistic when there are several loci affecting the trait for both dense and discrete markers.

A numerical evaluation of these formulas indicates that in the case of a single trait locus, the smoothed statistic has slightly less power than the original. This contradicts the suggestion (Hoh and Ott (2000) that power would increase because of the "true peaks are wider than false peaks" argument of Terwilliger *et al.* (1996). 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 statistics requires very little extra effort, we can say that in order to increase the chances to detect linkage it is worthwhile to consider the tests based on the scan statistic with different bandwidths. 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 several trait loci on a given chromosome.

Although the moving averge statistic appears to lose a small amount of power in comparison with the score statistic when there is only one linked gene, its power is not as 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 in 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 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 modest 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 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 nonlinear smoothing conducted at the level of the pedigree, not at the level of the statistic. 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 multipoint analysis, which will improve power to detect genes located near markers with low information content.

The method here described seems to be a useful tool in cases where the linked trait loci have a rather 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 an strong effect and is easy to detect, while the effect of the second one is relatively small and is masked by the major gene. We will analyze how a conditional approach for sequential detection of trait-loci performs in such situations in a future paper.

Appendix

(I) Variance of the scan statistic. Under the null hypothesis of no trait-loci, the process $D_{c,t}$ defined by (2) and (3) satisfies

$$\mathrm{E}_0\left[D_{c,t}\right] = \frac{1}{2} \quad \text{and} \quad \mathrm{Cov}_0\left[D_{c,t}, D_{c,s}\right] = \frac{e^{-\beta|t-s|}}{4},$$

and consequently for the score statistic $V_{c,t}$ defined in (6), $\mathrm{E}_0\left[V_{c,t}\right] = 0$ and $\mathrm{Cov}_0\left[V_{c,t}, V_{c,s}\right] = e^{-\beta|t-s|}$. This implies that the variance of the scan statistic $S^{\epsilon}_{c,t}$ defined in (7) is given by

$$\begin{aligned}
\mathrm{Var}_0\left[S^{\epsilon}_{c,t}\right] &= \mathrm{E}_0\left[\frac{1}{4\epsilon^2}\int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} V_{c,v}\, V_{c,u}\, du\, dv\right] \\
&= \frac{1}{4\epsilon^2}\int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} \mathrm{E}_0\left[V_{c,v}\, V_{c,u}\right]\, du\, dv \\
&= \frac{1}{4\epsilon^2}\int_{-\epsilon}^{\epsilon}\int_{-\epsilon}^{\epsilon} e^{-\beta|u-v|}\, du\, dv \\
&= \frac{2\beta\epsilon - 1 + e^{-2\beta\epsilon}}{2\beta^2\epsilon^2} \\
&= \left(2\epsilon\kappa^{\epsilon}\beta\right)^{-2}. \qquad (15)
\end{aligned}$$

Hence the standardized version of the scan statistic is

$$Q^{\epsilon}_{c,t} = \frac{S^{\epsilon}_{c,t}}{\sqrt{\mathrm{Var}_0\left[S^{\epsilon}_{c,t}\right]}} = \kappa^{\epsilon}\beta\int_{t-\epsilon}^{t+\epsilon} V_{c,s}\, ds,$$

as defined in (8).

(II) False positive error rates. It follows from the central limit theorem that, under the hypothesis of absence of trait-loci, as $N \to \infty$ the process V_t defined in (6) converges in distribution to a stationary Ornstein-Uhlenbeck process, Z_t, with mean value 0 and covariance function

$$\mathrm{Cov}[Z_t, Z_{t+s}] = e^{-\beta|s|}. \tag{16}$$

This implies that the standardized scan process Q_t^ϵ defined in (8) converges weakly to the smoothed process

$$X_t^\epsilon = \kappa^\epsilon \beta \int_{t-\epsilon}^{t+\epsilon} Z_s ds.$$

X_t^ϵ is a stationary Gaussian process with mean value 0 and covariance function, $R(s) = \mathrm{Cov}\left(X_t^\epsilon, X_{t+s}^\epsilon\right)$, given by

$$\begin{aligned} R(s) &= \left(\kappa^\epsilon\right)^2 \beta^2 \int_{-\epsilon}^{\epsilon} \int_{s-\epsilon}^{s+\epsilon} e^{-\beta|u-v|}\, du\, dv \qquad (17) \\ &= \begin{cases} \left(\kappa^\epsilon\right)^2 \left(e^{-\beta(2\epsilon-|s|)} + e^{-\beta(2\epsilon+|s|)} - 2e^{-\beta|s|} + 2(2\epsilon-|s|)\beta\right) & \text{if } |s| \le 2\epsilon, \\ \left(\kappa^\epsilon\right)^2 \left(e^{-\beta(|s|+2\epsilon)} + e^{-\beta(|s|-2\epsilon)} - 2e^{-\beta|s|}\right) & \text{if } |s| > 2\epsilon. \end{cases} \end{aligned}$$

The approximate formula (10) for false positive rate of the test based on Q_t^ϵ is an asymptotic approximation which comes from replacing the process Q_t^ϵ by X_t^ϵ and applying the following result.

Proposition 1. Let X_t^ϵ be a Gaussian process with mean 0 and covariance function (17). For $0 < \epsilon < L/2$ and large b,

$$P_0\left[\max_{\epsilon \le t \le L-\epsilon} X_t^\epsilon > b\right] \le 1 - \Phi(b) + \kappa^\epsilon \beta (L-2\epsilon)\phi(b)\sqrt{\frac{1-e^{-2\beta\epsilon}}{\pi}}.$$

Proof. The results in section 3 of Davies (1977), imply that

$$P_0\left[\max_{\epsilon \le t \le L-\epsilon} X_t^\epsilon > b\right] \le 1 - \Phi(b) + \frac{(L-2\epsilon)\sqrt{-R''(0)}\,\phi(b)}{\sqrt{2\pi}}.$$

By taking second derivative of (17) at 0 we get,

$$-R''(0) = 2\left(\kappa^\epsilon\right)^2 \beta^2 \left(1 - e^{-2\beta\epsilon}\right),$$

from which the result follows. □

(III) One trait-locus. Now we assume that the model with a unique locus trait at chromosome c_0 described in case (B) is valid for some locus position τ and $\alpha > 0$. Then, for $t \in c_0$, as $N \to \infty$ and $\alpha \to 0$ in such as way that $N^{1/2}\alpha \to \xi$, the scan process V_t converges to a Gaussian process, which we will denote again by Z_t. It has covariance function given by (16) and mean function

$$\mathrm{E}[Z_t] = \xi e^{-\beta|t-\tau|}. \tag{18}$$

Approximation (11) for the power of the score statistic is based on the process Z_t (see Feingold *et al.*, 1993 for details).

The limit in distribution of the scan process Q_t^ϵ is the Gaussian process $X_t^\epsilon = \kappa^\epsilon \beta \int_{t-\epsilon}^{t+\epsilon} Z_s ds$, with covariance function (17) and expectation function

$$\begin{aligned}\mu_\xi^\epsilon(t) &= \kappa^\epsilon \beta \xi \int_{t-\epsilon}^{t+\epsilon} e^{-\beta|s-\tau|}\, ds \qquad (19)\\ &= \begin{cases} \kappa^\epsilon \xi \left(e^{-\beta(\tau-\epsilon-t)} - e^{-\beta(\tau+\epsilon-t)}\right) & \text{if } t < \tau-\epsilon, \\ \kappa^\epsilon \xi \left(2 - e^{-\beta(\tau+\epsilon-t)} - e^{-\beta(t+\epsilon-\tau)}\right) & \text{if } \tau-\epsilon \le t \le \tau+\epsilon, \\ \kappa^\epsilon \xi \left(e^{-\beta(t-\epsilon-\tau)} - e^{-\beta(t+\epsilon-\tau)}\right) & \text{if } t > \tau+\epsilon. \end{cases}\end{aligned}$$

This function reaches its maximum at τ with maximum value

$$m_\xi^\epsilon = \mu_\xi^\epsilon(\tau) = 2\kappa^\epsilon \xi \left(1 - e^{-\beta\epsilon}\right). \qquad (20)$$

For large N, the power of the test based on the scan statistic, $P_\xi\left[\max_{\epsilon \le t \le L-\epsilon} Q_t^\epsilon \ge b\right]$, will be close to $P_\xi\left[\max_{\epsilon \le t \le L-\epsilon} X_t^\epsilon \ge b\right]$. The next proposition gives an approximate formula for this last probability, from which (12) follows.

PROPOSITION 2. Let X_t^ϵ be a Gaussian process with covariance function (17) and mean value (19). If $\epsilon < \tau < L - \epsilon$ then for $0 < \epsilon < L/2$ and large b and ξ the approximation

$$P_\xi\left[\max_{\epsilon \le t \le L-\epsilon} X_t^\epsilon \ge b\right] \simeq 1 - \Phi(b - m_\xi^\epsilon) + \frac{\phi(b - m_\xi^\epsilon)}{b - m_\xi^\epsilon}\left(\sqrt{1 + \frac{m_\xi^\epsilon (b - m_\xi^\epsilon)(1 + e^{\beta\epsilon})}{2\xi^2}} - 1\right),$$

holds with m_ξ^ϵ defined by (20).

PROOF. The target probability can be written as

$$\begin{aligned} & P_\xi\left[\max_{\epsilon \le t \le L-\epsilon} X_t^\epsilon \ge b\right] \\ &= P_\xi\left[X_\tau^\epsilon \ge b\right] + P_\xi\left[X_\tau^\epsilon < b,\ \max_t X_t^\epsilon \ge b\right] \\ &= 1 - \Phi(b - m_\xi^\epsilon) + \int_0^\infty P_\xi\left[\max_t X_t^\epsilon \ge b \,|\, X_\tau^\epsilon = b - s\right] \phi(b - s - m_\xi^\epsilon) ds. \qquad (21)\end{aligned}$$

We approximate the integral by an approach similar to that in Section 5 of Siegmund and Worsley (1995). They reason that, conditional on $X_\tau^\epsilon = b - s$ with $s \approx 0$, for $t - \tau \simeq 0$ the process X_t^ϵ satisfies

$$X_t^\epsilon \approx b - s + (t-\tau)\dot{X}_\tau^\epsilon + \frac{(t-\tau)^2}{2} \mathrm{E}\left[\ddot{X}_\tau^\epsilon \,|\, X_\tau^\epsilon = b - s\right], \qquad (22)$$

and a maximization of the right-hand side of (22) yields

$$\max_t X_t^\epsilon \approx b - s - \frac{\left(\dot{X}_\tau^\epsilon\right)^2}{2\mathrm{E}\left[\ddot{X}_\tau^\epsilon \,|\, X_\tau^\epsilon = b - s\right]}. \qquad (23)$$

Let $g^\epsilon = -R''(0) = 2\left(\kappa^\epsilon\right)^2 \beta^2 (1 - e^{-2\beta\epsilon})$. For our process, $\mathrm{E}[\ddot{X}_\tau^\epsilon] = \mu_\xi^{\epsilon\prime\prime}(\tau) = -2\kappa^\epsilon \beta^2 e^{-\beta\epsilon}\xi$, and $\mathrm{Cov}(X_\tau^\epsilon, \ddot{X}_\tau^\epsilon) = R''(0) = -g^\epsilon$, and hence by joint normality

$$\begin{aligned}\mathrm{E}\left[\ddot{X}_\tau^\epsilon \,|\, X_\tau^\epsilon = b - s\right] &= \mathrm{E}\left[\ddot{X}_\tau^\epsilon\right] + \mathrm{Cov}\left(X_\tau^\epsilon, \ddot{X}_\tau^\epsilon\right)\left(b - s - m_\xi^\epsilon\right) \\ &= -\left[2\,\kappa^\epsilon \beta^2 e^{-\beta\epsilon}\xi + g^\epsilon\left(b - s - m_\xi^\epsilon\right)\right] \\ &\simeq -\left[2\,\kappa^\epsilon \beta^2 e^{-\beta\epsilon}\xi + g^\epsilon\left(b - m_\xi^\epsilon\right)\right]. \qquad (24)\end{aligned}$$

The last approximation comes from the fact that we are assuming that b is large and the relevant values of s are close to 0. Let $l_\xi^\epsilon = 2\,\kappa^\epsilon \beta^2 e^{-\beta\epsilon}\xi + g^\epsilon(b - m_\xi^\epsilon)$ and $w_\xi^\epsilon = 2l_\xi^\epsilon/g^\epsilon$. Because of (23) we have

$$\max_t X_t^\epsilon \simeq b - s + \frac{\left(\dot{X}_\tau^\epsilon\right)^2}{2l_\xi^\epsilon},$$

and thus

$$P_\xi\left[\max_{\epsilon\le t\le L-\epsilon} X_t^\epsilon \ge b \,|\, X_\tau^\epsilon = b - s\right] \simeq P_\xi\left[\frac{\left(\dot{X}_\tau^\epsilon\right)^2}{2l_\xi^\epsilon} \ge s\right] = P_\xi\left[\frac{\left(\dot{X}_\tau^\epsilon\right)^2}{g^\epsilon} \ge \frac{2l_\xi^\epsilon}{g^\epsilon}s\right] = P\left[\chi_1^2 \ge w_\xi^\epsilon s\right].$$

If we neglect the term involving s^2 in $\phi(b - s - m_\xi^\epsilon)$, we get

$$\begin{aligned}
&\int_0^\infty P_\xi\left[\max_{\epsilon\le t\le L-\epsilon} X_t^\epsilon \ge b \,|\, X_\tau = b - s\right]\phi(b - s - m_\xi^\epsilon)ds \\
&\simeq \int_0^\infty P\left[\chi_1^2 \ge w_\xi^\epsilon s\right]\phi(b - s - m_\xi^\epsilon)\,ds \\
&\simeq \phi(b - m_\xi^\epsilon)\int_0^\infty P\left[\chi_1^2 \ge w_\xi^\epsilon s\right]\exp\left((b - m_\xi^\epsilon)s\right)\,ds \\
&= \phi(b - m_\xi^\epsilon)\int_0^\infty \exp\left((b - m_\xi^\epsilon)s\right)\left(\int_{w_\xi^\epsilon s}^\infty \frac{1}{\sqrt{2\pi}}\,y^{-1/2}\,e^{-y/2}dy\right)ds \\
&= \phi(b - m_\xi^\epsilon)\int_0^\infty \frac{1}{\sqrt{2\pi}}\,y^{-1/2}\,e^{-y/2}\left(\int_0^{y/w_\xi^\epsilon}\exp\left((b - m_\xi^\epsilon)s\right)ds\right)dy \\
&= \frac{\phi(b - m_\xi^\epsilon)}{b - m_\xi^\epsilon}\left(\int_0^\infty \frac{1}{\sqrt{2\pi}}\,y^{-1/2}\exp\left(-y\frac{w_\xi^\epsilon + 2(m_\xi^\epsilon - b)}{2w_\xi^\epsilon}\right)dy - 1\right) \\
&= \frac{\phi(b - m_\xi^\epsilon)}{b - m_\xi^\epsilon}\left(\sqrt{\frac{w_\xi^\epsilon}{w_\xi^\epsilon + 2(m_\xi^\epsilon - b)}} - 1\right).
\end{aligned}$$

Combining this with (21) we obtain the proposition. □

(IV) Two trait-loci on the same chromosome. Finally we will analyse the model described in (C), where a chromosome c_0 of genetic length L contains two trait-loci at positions $0 < \tau_1 < \tau_2 < L$ with additive effects $\alpha_1, \alpha_2 > 0$. Let $d = |\tau_2 - \tau_1|$ denote the distance between the two trait-loci, $\rho = e^{-\beta d}$ the correlation between Z_{τ_1} and Z_{τ_2}, $\xi_i = \sqrt{N}\,\alpha_i$ for $i = 1, 2$, and h the density function of a bivariate normal distribution

with mean vector $(\xi_1 + \rho\,\xi_2, \rho\,\xi_1 + \xi_2)'$, variances equal to 1, and correlation ρ, that is,

$$h(x,y) = \frac{1}{\sqrt{1-\rho^2}}\,\phi(x - \xi_1 - \rho\,\xi_2)\phi\left(\frac{y - \rho\,\xi_1 - \xi_2 - \rho\,(x - \xi_1 - \rho\xi_2)}{\sqrt{1-\rho^2}}\right),$$

We assume, without loss of generality, that $\xi_1 > \xi_2$.

[(IV) - 1] Power of the test based on the score statistic. Under this model V_t converges in distribution to a Gaussian process, Z_t, with covariance function given by (16) and mean function

$$\mathrm{E}[Z_t] = \xi_1 e^{-\beta|t-\tau_1|} + \xi_2 e^{-\beta|t-\tau_2|}. \tag{25}$$

For large N the power to detect linkage to chromosome c_0 of the test based on the scan statistic will be close to the probability that the maximum of the process Z_t over the interval $[0, L]$ exceeds the threshold b.

Proposition 3. Let Z_t be a Gaussian process with covariance function (16) and mean function (25). For $L > \tau_2$, and large values of b, ξ_1 and ξ_2,

$$\begin{aligned} & P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b\right] \\ & \quad\simeq\quad 1 - \int_0^\infty \Phi\left(\frac{b - \rho(b-x) - \xi_2(1-\rho^2)}{\sqrt{1-\rho^2}}\right)\phi\left(b - x - \xi_1 - \rho\xi_2\right)dx \\ & \qquad + \int_0^\infty\int_0^\infty q(x,y)\,h(b-x, b-y)dx\,dy, \end{aligned}$$

where

$$q(x,y) = \left(e^{-bx} + e^{-by} - e^{-b(x+y)}\right)(1 - q_c(x,y)) + q_c(x,y),$$

with

$$q_c(x,y) = p_c(x) + p_c(y) - p_c(x)p_c(y),$$

and

$$p_c(x) = 1 - \Phi\left(\frac{b(1-\rho)\sqrt{\beta d}}{2(1+\rho)} + \frac{x}{\sqrt{\beta d}}\right) + \Phi\left(\frac{b(1-\rho)\sqrt{\beta d}}{2(1+\rho)} - \frac{x}{\sqrt{\beta d}}\right)\,e^{-bx(1-\rho)/(1+\rho)}.$$

Proof. The target probability can be written as

$$\begin{aligned} & P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b\right] \qquad (26)\\ & \quad= \quad P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max\left\{Z_{\tau_1}, Z_{\tau_2}\right\} > b\right] + P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b, \max\left\{Z_{\tau_1}, Z_{\tau_2}\right\} \le b\right]. \end{aligned}$$

For the first term in the right-hand side of (26) we have the identities

$$\begin{aligned} & P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max\left\{Z_{\tau_1}, Z_{\tau_2}\right\} > b\right] \\ & \quad= \quad 1 - P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[Z_{\tau_1} \le b, Z_{\tau_2} \le b\right] \\ & \quad= \quad 1 - \int_0^\infty P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[Z_{\tau_2} \le b \,|\, Z_{\tau_1} = b - x\right]\phi\left(b - x - \xi_1 - \rho\xi_2\right)dx \\ & \quad= \quad 1 - \int_0^\infty \Phi\left(\frac{b - \rho(b-x) - \xi_2(1-\rho^2)}{\sqrt{1-\rho^2}}\right)\phi\left(b - x - \xi_1 - \rho\xi_2\right)dx. \qquad (27) \end{aligned}$$

On the other hand, the second term in the right-hand side of (26) can be expressed as

$$P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b, \max\{Z_{\tau_1}, Z_{\tau_2}\} \le b\right] \tag{28}$$
$$= \int_0^\infty \int_0^\infty P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right] h(b-x, b-y)dx\,dy.$$

Since the events $\{\max_{t<\tau_1} Z_t > b\}$, $\{\max_{t>\tau_2} Z_t > b\}$ and $\{\max_{\tau_1<t<\tau_2} Z_t > b\}$ are conditionally independent given Z_{τ_1} and Z_{τ_2}, the conditional probability in the integral at the right-hand side of (28) can be expressed, for each value of x and y, as

$$P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0\le t\le L} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right]$$
$$= (1-q_d(x,y))\,(q_l(x,y) + q_h(x,y) - q_l(x,y)q_h(x,y)) + q_d(x,y), \tag{29}$$

where

$$q_l(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t<\tau_1} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right],$$

$$q_h(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t>\tau_2} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right],$$

and

$$q_d(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\tau_1<t<\tau_2} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right].$$

The conditional probability

$$P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t<\tau_1} Z_t > b \,|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right] = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t<\tau_1} Z_t > b \mid Z_{\tau_1} = b-x\right]$$

is negligible unless x is close to 0. For $x \approx 0$ and $t < \tau_1$ with $(t-\tau_1) \approx 0$,

$$\mathrm{E}\left[Z_t - Z_{\tau_1} \,|\, Z_{\tau_1} = b-x\right] = \left(1 - e^{-\beta(\tau_1 - t)}\right)(b-x) \simeq -\beta b(\tau_1 - t),$$

and

$$\mathrm{Var}\left[Z_t \,|\, Z_{\tau_1} = b-x\right] = 1 - e^{-2\beta(\tau_1-t)} \simeq 2\beta(\tau_1 - t).$$

Hence, conditional on $Z_{\tau_1} = b-x$ with x small, when t is smaller than τ_1 and close to it, the process Z_t behaves like a Brownian motion, W_t, with drift coefficient $-\beta b$ and diffusion coefficient 2β. Consequently,

$$\begin{aligned} q_l(x,y) &= P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t<\tau_1} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right] \\ &= P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t<\tau_1} Z_t > b \,|\, Z_{\tau_1} = b-x\right] \\ &\simeq P\left[\max_{t<0} W_t > x\right] \\ &= e^{-bx}. \end{aligned} \tag{30}$$

A similar argument leads to the approximation

$$q_h(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{t>\tau_2} Z_t > b \,\middle|\, Z_{\tau_1} = b-x, Z_{\tau_2} = b-y\right] \simeq e^{-by}. \tag{31}$$

Furthermore, for $\tau_1 < t < \tau_2$,

$$\begin{aligned}&\mathrm{E}\left[Z_t \,|Z_{\tau_1} = b - x, Z_{\tau_2} = b - y\right] \\ &= \frac{b\left(e^{-\beta(t-\tau_1)} + e^{-\beta(\tau_2 - t)}\right)}{1+\rho} - \frac{e^{-\beta(t-\tau_1)}(x - \rho y) + e^{-\beta(\tau_2-t)}(y - \rho x)}{1-\rho^2},\end{aligned}$$

and

$$\mathrm{Var}\left[Z_t \,|Z_{\tau_1} = b - x, Z_{\tau_2} = b - y\right] = 1 - \frac{e^{-2\beta(t-\tau_1)} + e^{-2\beta(\tau_2 - t)} - 2\rho^2}{1-\rho^2}.$$

In order to compute $q_d(x, y)$ we will divide the interval (τ_1, τ_2) into two halves. Let us define

$$q_{d1}(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\tau_1 < t < (\tau_1+\tau_2)/2} Z_t > b \,\middle|\, Z_{\tau_1} = b - x, Z_{\tau_1} = b - y\right]$$

and

$$q_{d2}(x,y) = P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{(\tau_1+\tau_2)/2 < t < \tau_2} Z_t > b \,\middle|\, Z_{\tau_1} = b - x, Z_{\tau_1} = b - y\right].$$

The events $\left[\max_{(\tau_1 < t < (\tau_1+\tau_2)/2} Z_t > b\right]$ and $\left[\max_{(\tau_1+\tau_2)/2 < t < \tau_2} Z_t > b\right]$ are asymptotically conditionally independent given Z_{τ_1} and Z_{τ_2}, and thus

$$q_d(x,y) \simeq q_{d1}(x,y) + q_{d2}(x,y) - q_{d1}(x,y) q_{d2}(x,y). \tag{32}$$

We will restrict attention to the cases where x and y are close to 0, since the conditional probability $P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\tau_1 < t < \tau_2} Z_t > b \,|\, Z_{\tau_1} = b - x, Z_{\tau_2} = b - y\right]$ is negligible otherwise. If $x \approx 0$ and $y \approx 0$, then for $t > \tau_1$ and $(t - \tau_1)$ small,

$$\mathrm{E}\left[\, Z_t - Z_{\tau_1} \,|\, Z_{\tau_1} = b - x, Z_{\tau_2} = b - y\right] \simeq -\beta b (t - \tau_1)\frac{1-\rho}{1+\rho},$$

and

$$\mathrm{Var}\left[Z_t \,|Z_{\tau_1} = b - x\right] \simeq 2\beta(t - \tau_1).$$

Hence we can use the approximation

$$q_{d1} \simeq P\left[\max_{0 < t < d/2} B_t > x\right] = p_c(x), \tag{33}$$

say, where B_t denotes a Brownian motion with drift coefficient $-\beta b(1-\rho)/(1+\rho)$ and diffusion coefficient 2β (see for example (3.15) in Siegmund, 1985). By symmetry, for $t < \tau_2$ and $(\tau_2 - t) \approx 0$ the approximation

$$q_{d2}(x,y) = p_c(y) \tag{34}$$

holds. The final approximate formula for the probability of exceeding b in the interval $[0, L]$ is obtained by substituting (33) and (34) in (32); (30), (31) and (32) in (29), and (27) and (29) in (26). □

Suppose now that we look at the discrete set of random variables $Z_{i\Delta}$, for $0 \le i\Delta \le L$. Proposition 3* gives an approximate formula for

$$P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{0 \le i\Delta \le L} Z_t > b\right].$$

PROPOSITION 3*. Let Z_t be the Gaussian process defined in Proposition 3. If $b \to \infty$, $\Delta \to 0$ in such a way that $b\sqrt{\Delta}$ converges to a finite constant, then

$$\begin{aligned} P\left[\max_{0\le i\Delta\le L} Z_t > b\right] &\simeq 1 - \int_0^\infty \Phi\left(\frac{b-\rho(b-x)-\xi_2(1-\rho^2)}{\sqrt{1-\rho^2}}\right)\phi\left(b-x-\xi_1-\rho\xi_2\right)dx \\ &\quad + \int_0^\infty \int_0^\infty q^\Delta(x,y)\, h(b-x,b-y)dx\,dy, \end{aligned}$$

where

$$\begin{aligned} &q^\Delta(x,y) \\ &= \left(e^{-b\left(x+\varrho\sqrt{2\beta\Delta}\right)} + e^{-b\left(y+\varrho\sqrt{2\beta\Delta}\right)} - e^{-b\left(x+y+2\varrho\sqrt{2\beta\Delta}\right)}\right)\left(1-q_c^\Delta(x,y)\right) + q_c^\Delta(x,y), \end{aligned}$$

with

$$q_c^\Delta(x,y) = p_c^\Delta(x) + p_c^\Delta(y) - p_c^\Delta(x)\, p_c^\Delta(y),$$

and

$$\begin{aligned} p_c^\Delta(x) &= 1 - \Phi\left(\frac{b(1-\rho)\sqrt{\beta d}}{2(1+\rho)} + \frac{x}{\sqrt{\beta d}}\right) \\ &\quad + \Phi\left(\frac{b(1-\rho)\sqrt{\beta d}}{2(1+\rho)} - \frac{x+2\varrho\sqrt{2\beta\Delta}}{\sqrt{\beta d}}\right) e^{-b(x+\varrho\sqrt{2\beta\Delta})(1-\rho)/(1+\rho)}, \end{aligned}$$

where ϱ is the constant defined in (3.29) of Siegmund (1985).

PROOF. To derive this result we follow similar steps as in the proof of Proposition 3. The necessary changes are to enclose a factor $\exp\left(-b\varrho\sqrt{2\beta\Delta}\right)$ in the approximate expressions for $q_l(x,y)$ and $q_h(x,y)$ (see Appendix A of Feingold et al, 1993 for the details), and to replace (3.15) of Siegmund (1985) by (3.28)-(3.29) of the same book in the approximate computation of $q_{d1}(x,y)$ and $q_{d2}(x,y)$. □

[(IV) - 2] POWER OF THE TEST BASED ON THE SCAN STATISTIC. The limit in distribution of the scan statistic, $X_t^\epsilon = \kappa^\epsilon\beta\int_{t-\epsilon}^{t+\epsilon} Z_s ds$, is under this situation a Gaussian process with covariance function given by (17) and its expectation function,

$$\mu^\epsilon(t) = \mathrm{E}_{\xi_1,\xi_2,\tau_1,\tau_2}\left[X_t^\epsilon\right],$$

depends on relationships between all parameters in a complicated way. For $\epsilon \ge \tau_2 - \tau_1$,

$$\mu^\epsilon(t) =$$

$$\begin{cases} \kappa^\epsilon\left[\xi_1\left(e^{-\beta(\tau_1-t-\epsilon)} - e^{-\beta(\tau_1-t+\epsilon)}\right) + \xi_2\left(e^{-\beta(\tau_2-t-\epsilon)} - e^{-\beta(\tau_2-t+\epsilon)}\right)\right] & \text{if } t < \tau_1-\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(2 - e^{-\beta(\tau_1-t+\epsilon)} - e^{-\beta(t+\epsilon-\tau_1)}\right) + \xi_2\left(e^{-\beta(\tau_2-t-\epsilon)} - e^{-\beta(\tau_2-t+\epsilon)}\right)\right] & \text{if } \tau_1-\epsilon \le t < \tau_2-\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(2 - e^{-\beta(\tau_1-t+\epsilon)} - e^{-\beta(t+\epsilon-\tau_1)}\right) + \xi_2\left(2 - e^{-\beta(\tau_2-t+\epsilon)} - e^{-\beta(t+\epsilon-\tau_2)}\right)\right] & \text{if } \tau_2-\epsilon \le t < \tau_1+\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(e^{-\beta(t-\epsilon-\tau_1)} - e^{-\beta(t+\epsilon-\tau_1)}\right) + \xi_2\left(2 - e^{-\beta(\tau_2-t+\epsilon)} - e^{-\beta(t+\epsilon-\tau_2)}\right)\right] & \text{if } \tau_1+\epsilon \le t < \tau_2+\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(e^{-\beta(t-\epsilon-\tau_1)} - e^{-\beta(t+\epsilon-\tau_1)}\right) + \xi_2\left(e^{-\beta(t-\epsilon-\tau_2)} - e^{-\beta(t+\epsilon-\tau_2)}\right)\right] & \text{if } t \ge \tau_2+\epsilon, \end{cases}$$

holds, and for $\epsilon < \tau_2 - \tau_1$

$$\mu^\epsilon(t) =$$

$$\begin{cases} \kappa^\epsilon\left[\xi_1\left(e^{-\beta(\tau_1-t-\epsilon)}-e^{-\beta(\tau_1-t+\epsilon)}\right)+\xi_2\left(e^{-\beta(\tau_2-t-\epsilon)}-e^{-\beta(\tau_2-t+\epsilon)}\right)\right] & \text{if } t<\tau_1-\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(2-e^{-\beta(\tau_1-t+\epsilon)}-e^{-\beta(t+\epsilon-\tau_1)}\right)+\xi_2\left(e^{-\beta(\tau_2-t-\epsilon)}-e^{-\beta(\tau_2-t+\epsilon)}\right)\right] & \text{if } \tau_1-\epsilon\le t<\tau_1+\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(e^{-\beta(t-\epsilon-\tau_1)}-e^{-\beta(t+\epsilon-\tau_1)}\right)+\xi_2\left(e^{-\beta(\tau_2-t-\epsilon)}-e^{-\beta(\tau_2-t+\epsilon)}\right)\right] & \text{if } \tau_1+\epsilon\le t<\tau_2-\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(e^{-\beta(t-\epsilon-\tau_1)}-e^{-\beta(t+\epsilon-\tau_1)}\right)+\xi_2\left(2-e^{-\beta(\tau_2-t+\epsilon)}-e^{-\beta(t+\epsilon-\tau_2)}\right)\right] & \text{if } \tau_2-\epsilon\le t<\tau_2+\epsilon, \\ \kappa^\epsilon\left[\xi_1\left(e^{-\beta(t-\epsilon-\tau_1)}-e^{-\beta(t+\epsilon-\tau_1)}\right)+\xi_2\left(e^{-\beta(t-\epsilon-\tau_2)}-e^{-\beta(t+\epsilon-\tau_2)}\right)\right] & \text{if } t\ge\tau_2+\epsilon. \end{cases}$$

Let

$$\begin{aligned} t^\epsilon &= t^\epsilon(\xi_1,\xi_2,\tau_1,\tau_2)=\arg\max_{\epsilon\le t\le L-\epsilon}\mu^\epsilon(t), \\ m^\epsilon &= m^\epsilon(\xi_1,\xi_2,\tau_1,\tau_2)=\max_{\epsilon\le t\le L-\epsilon}\mu^\epsilon(t)=\mu^\epsilon\left(t^\epsilon\right), \\ a^* &= a^*(\xi_1,\xi_2,\tau_1,\tau_2)=\frac{1}{2\beta}\log\left(\frac{\xi_1+\xi_2/\rho}{\xi_1+\xi_2\rho}\right), \\ b^* &= b^{\epsilon *}(\xi_1,\xi_2,\tau_1,\tau_2)=\frac{1}{2\beta}\log\left(\frac{\xi_1}{\xi_1+\rho\xi_2(1-e^{2\beta\epsilon})}\right), \end{aligned}$$

and

$$\epsilon^*=\epsilon^*(\xi_1,\xi_2,\tau_1,\tau_2)=\frac{d}{2}-\frac{1}{2\beta}\log\left(\frac{\xi_1}{\xi_2}\right). \tag{35}$$

Note that a^* does not depends on ϵ. By using some rather elementary but tiresome algebra we find

$$t^\epsilon=\begin{cases}\tau_1+a^* & \text{if } \epsilon\ge d-a^*, \\ \tau_1+b^* & \text{if } \epsilon<d-a^*,\end{cases} \tag{36}$$

and

$$m^\epsilon=\begin{cases} 2\kappa^\epsilon\left(\xi_1+\xi_2-e^{-\beta\epsilon}\sqrt{\xi_1^2+\xi_2^2+\xi_1\xi_2(\rho+1/\rho)}\right) & \text{if } \epsilon\ge d-a^*, \\ 2\kappa^\epsilon\left(\xi_1-e^{-\beta\epsilon}\sqrt{\xi_1^2+\rho\xi_1\xi_2\left(1-e^{2\beta\epsilon}\right)}\right) & \text{if } \epsilon<d-a^*, \end{cases} \tag{37}$$

and that $\mu^\epsilon(t)$ has no other local maximum provided that $\epsilon\ge\epsilon^*$. The next result gives an approximate formula for the probability that the maximum of X_t^ϵ exceeds b for this case.

Proposition 4. Let X_t^ϵ be a Gaussian process with covariance function (17) and mean function $\mu^\epsilon(t)$. If $\epsilon<\tau_1<\tau_2<L-\epsilon$, then for $\epsilon\ge\epsilon^*$ and large values of b, ξ_1 and ξ_2,

$$P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\epsilon\le t\le L-\epsilon}X_t^\epsilon\ge b\right]$$

$$\simeq\begin{cases} 1-\Phi(b-m^\epsilon)+\frac{\phi(b-m^\epsilon)}{b-m^\epsilon}\left(\sqrt{1+\frac{2(\kappa^\epsilon)^2(b-m^\epsilon)(1-e^{-2\beta\epsilon})}{2\kappa^\epsilon(\xi_1+\xi_2)-m^\epsilon}}-1\right) & \text{if } \epsilon\ge d-a^*, \\ 1-\Phi(b-m^\epsilon)+\frac{\phi(b-m^\epsilon)}{b-m^\epsilon}\left(\sqrt{1+\frac{2(\kappa^\epsilon)^2(b-m^\epsilon)(1-e^{-2\beta\epsilon})}{2\kappa^\epsilon\xi_1-m^\epsilon}}-1\right) & \text{if } d-a^*<\epsilon<\epsilon^*, \end{cases}$$

with m^ϵ defined by (37).

PROOF. Since $\mu^\epsilon(t)$ has only one maximum when $\epsilon > \epsilon^*$, an approach similar to that on Proposition 2 can be used in such cases. If $\epsilon \geq d - a^*$, then $t^\epsilon = \tau_1 + a^*$,

$$m^\epsilon = 2\kappa^\epsilon \left(\xi_1 + \xi_2 - e^{-\beta\epsilon} \sqrt{\xi_1^2 + \xi_2^2 + \xi_1\xi_2(\rho + 1/\rho)} \right), \tag{38}$$

and

$$\mathrm{E}\left[\ddot{X}_{t^\epsilon}\right] = \mu^{\epsilon\prime\prime}(\tau_1 + a^*) = -2\kappa^\epsilon\beta^2 e^{-\beta\epsilon}\sqrt{\xi_1^2 + \xi_2^2 + \xi_1\xi_2(\rho + 1/\rho)}.$$

By following the same steps as in the proof of Proposition 2 we get

$$\begin{aligned}
&P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\epsilon\leq t\leq L-\epsilon} X_t^\epsilon \geq b\right] \\
&\quad = \quad P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[X^\epsilon_{\tau_1+a^*} \geq b\right] + P_{\xi_1,\xi_2,\tau_1,\tau_2}\ \left[X^\epsilon_{\tau_1+a^*} < b,\ \max_t X_t^\epsilon \geq b\right] \\
&\quad \simeq \quad 1 - \Phi(b - m^\epsilon) + \frac{\phi(b-m^\epsilon)}{b-m^\epsilon}\left(\sqrt{1 + \frac{2(\kappa^\epsilon)^2(b-m^\epsilon)(1-e^{-2\beta\epsilon})}{2\kappa^\epsilon(\xi_1+\xi_2) - m^\epsilon}} - 1\right),
\end{aligned}$$

with m^ϵ given by (38).

On the other hand, when $\epsilon^* < \epsilon < d - a^*$, $\mu^\epsilon(t)$ attains its only maximum at $t^\epsilon = \tau_1 + b^*$, and hence

$$m^\epsilon = 2\kappa^\epsilon \left(\xi_1 - e^{-\beta\epsilon} \sqrt{\xi_1^2 + \rho\xi_1\xi_2\left(1 - e^{2\beta\epsilon}\right)} \right), \tag{39}$$

and

$$\mathrm{E}\left[\ddot{X}_{t^\epsilon}\right] = \mu^{\epsilon\prime\prime}(\tau_1 + b^*) = -2\kappa^\epsilon\beta^2 e^{-\beta\epsilon}\sqrt{\xi_1^2\rho\xi_1\xi_2\left(1 - e^{2\beta\epsilon}\right)},$$

which leads to the approximation

$$\begin{aligned}
&P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[\max_{\epsilon\leq t\leq L-\epsilon} X_t^\epsilon \geq b\right] \\
&\quad = \quad P_{\xi_1,\xi_2,\tau_1,\tau_2}\left[X^\epsilon_{\tau_1+b^*} \geq b\right] + P_{\xi_1,\xi_2,\tau_1,\tau_2}\ \left[X^\epsilon_{\tau_1+b^*} < b,\ \max_t X_t^\epsilon \geq b\right] \\
&\quad \simeq \quad 1 - \Phi(b - m^\epsilon) + \frac{\phi(b-m^\epsilon)}{b-m^\epsilon}\left(\sqrt{1 + \frac{2\left(\kappa^\epsilon\right)^2(b-m^\epsilon)(1-e^{-2\beta\epsilon})}{2\kappa^\epsilon\xi_1 - m^\epsilon}} - 1\right),
\end{aligned}$$

with m^ϵ defined by (39). □

Remark. When $\epsilon \leq \epsilon^*$, the expectation function $\mu^\epsilon(t)$ reaches its maximum value at $t^\epsilon = \tau_1 + b^*$, but has a second local maximum at $\tau_2 - c^*$, where

$$c^* = c^*_{\xi_1,\xi_2,\tau_1,\tau_2,\epsilon} = \frac{1}{2\beta}\ \log\left(\frac{\xi_2}{\xi_1\rho(1 - e^{2\beta\epsilon}) + \xi_2}\right).$$

In this case an approximation based on a unique maximum –as the one used in Proposition 4– is not appropriate.

Acknowledgments. The authors are grateful to Mathisca de Gunst, Chris Klaassen and Dorret Boomsma for helpful discussion and encouragement.

References

Cox D.R, Hinkley D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Davies R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–254.

Dupuis J, Brown P, Siegmund D. (1993). 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. *Genet. Epidemiol.*, **14**, 103–115.

Feingold E, Brown P, Siegmund D. (1993). Gaussian Models for Genetic Linkage Analysis Using Complete High-Resolution Maps of Identity by Descent. *Am. J. Hum. Genet.,* **53**, 234-251.

Hoh J, Ott J. (2000). Scan statistics to scan markers for susceptibility genes. *Proc. Natl. Acad. Sci.,* **95**, 9615–9617.

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

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

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

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*.

Siegmund D, 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, Read A.P. (1996) *Human Moleuclar Genetics*, BIOS Scientific Publishers, Oxford.

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, *Am. J. Hum. Genet.* **61**, 430-438.