Genetic Epidemiology 25: 59–72 (2003)

Assessing Accuracy in Linkage Analysis by Means of Confidence Regions

Ola Hössjer*

Department of Mathematics, Stockholm University, Stockholm, Sweden

When statistical linkage to a certain chromosomal region has been found, it is of interest to develop methods quantifying the accuracy with which the disease locus can be mapped. In this paper, we investigate the performance of three different types of confidence regions, with asymptotically correct coverage probability as the number of pedigrees grows. Our setup is that of a saturated map of marker data. We allow for arbitrary combinations of pedigree structures, and treat various kinds of genetic models (e.g. binary and quantitative phenotypes) in a unified way. The linkage scores are weighted sums of the individual family scores, with NPL and lod scores as special cases. We show that the expected length of the confidence region is inversely proportional to the slope-to-noise ratio, or equivalently, inversely proportional to the product of the square of the noncentrality parameter and a certain normalized slope-to-noise ratio. Our investigations reveal that maximal expected linkage scores can be quite different from estimation-based performance criteria based on expected length of confidence regions. The main reason is that there is no simple relationship between peak height and peak slope of the mean linkage score. One application of our results is planning of linkage studies: given a certain genetic model, we can approximate the number of pedigrees needed to obtain a confidence region with given coverage probability and expected length. *Genet Epidemiol* 25:59–72, 2003. © 2003 Wiley-Liss, Inc.

Key words: confidence region; linkage analysis; locus estimation; perfect marker information; slope-to-noise ratio

Grant sponsor: Swedish Research Council; Grant number: 629-2002-6286.
*Correspondence to: Ola Hössjer, Department of Mathematics, Stockholm University, S-106 91 Stockholm, Sweden.
E-mail: ola@math.su.se
Received for publication 16 June 2002; Revision accepted 28 January 2003
Published online in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.10248

INTRODUCTION

The goal of statistical linkage analysis is to test if any disease susceptibility genes are located on one or several chromosomal regions, and then to decrease as much as possible the size of the region showing statistical significance for harboring the disease gene. This is often a two-step procedure. Once statistical linkage to a region is found, it is of interest to know how well the disease locus can be estimated. For instance, suppose that the linkage study is followed by fine mapping using association analysis. Then the multiple testing problem associated with the latter is reduced, if search for the gene is narrowed down to a smaller region.

The traditional performance criterion in linkage analysis is the power to detect linkage. The most common setup of a test is to reject the null hypothesis of no linkage when the maximum linkage score exceeds a certain threshold. Feingold et al. [1993] showed that the power to detect linkage can be approximated by an explicit formula for large sample sizes. The leading term of this formula depends on the chosen significance level of the test and the noncentrality parameter η (the mean of the linkage score at the disease locus). Hence η (which is a pointwise criterion evaluated at the disease locus) is closely related to the power to detect linkage (which is a genomewide criterion). In fact, η was used as performance criterion by Sham et al. [1997] and Nilsson [1999]. Its main advantages are simplicity of computation and no need for specification of significance level. Relatively little work has been done on disease locus estimation, although the scientific problem of interest is a location-estimation problem. New work on confidence regions is therefore very important.

Classical parametric linkage analysis is based on the so-called lod score, which is the base 10 logarithm of the likelihood ratio for the hypothesis test that the disease locus is linked to a given

chromosomal position. The lod score is then maximized over the chromosomal regions of interest. The position attaining the maximum lod score coincides with the maximum likelihood estimator of the disease locus. If a fixed number of markers is used, it can be shown that the disease locus estimator converges at a rate $N^{-\frac{1}{2}}$ towards the true disease locus when N pedigrees are used [Ott, 1999]. However, if the assumed model is incorrectly specified, the disease locus estimator can in fact be inconsistent [Clerget-Darpoux et al., 1986]. A more robust disease locus estimator converging at rate $N^{-\frac{1}{2}}$ was recently proposed by Liang et al. [2001], using generalized estimating equations.

In view of the current availability of high-density maps of single-nucleotide polymorphisms (SNPs), it is of interest to study the behavior of linkage procedures when the markers give perfect inheritance information at all loci. Some authors have noticed that the size of confidence regions for the disease locus decreases at a faster rate N^{-1} under perfect marker information. Kong and Wright [1994] established the limiting distribution of the disease locus estimator for backcross designs. Darvasi et al. [1993] and Darvasi and Soller [1997] showed by simulation that the lengths of confidence intervals are inversely proportional to the sample size for backcross and F_2 designs. Dupuis and Siegmund [1999] gave theoretical justification of their results, using an asymptotic expansion of the expected length of the confidence interval. Kruglyak and Lander [1995] gave analytical expressions for the distribution function of confidence region lengths for affected-relative pairs in nonparametric linkage (NPL). Hössjer [2001a] recently established N^{-1} convergence and the limiting distributions for the disease locus estimator, i.e., the chromosomal position maximizing the linkage score. This result was based on defining linkage family scores conditionally on observed phenotypes (dichotomous or quantitative), and it holds for arbitrary pedigree structures, score functions, and weighting schemes.

The purpose of this paper is to use the results in Hössjer [2001a] to define confidence regions with asymptotically valid coverage probabilities and expected lengths. We define three different types of confidence regions: the support region, the convex support region, and an estimation-based confidence region. Their performance is investigated by simulation for affected relative pairs. It turns out that the asymptotic approximations are accurate for noncentrality parameters η of about 4 or larger, or for expected confidence region lengths about 5 centiMorgans (cM) or shorter.

One consequence of this work is to use the expected length of the confidence region as an alternative performance criterion. Such a criterion is not identical to the noncentrality parameter, because the expected confidence region length also depends on the slope of the mean linkage score at the disease locus and the amount of random fluctuation of the linkage score around the disease locus. The expected length of the confidence region is inversely proportional to the product of η^2 and a certain normalized slope-to-noise ratio C. Hence, C contains the residual "estimation information" present in the data set when the noncentrality parameter has been accounted for. It turns out that C varies a lot between data sets, genetic models, and chosen score functions. This strongly indicates that the estimation and testing criteria in linkage analysis are quite different, at least when the maximum linkage score is used as test statistic.

We refer to a genetic model as strong or weak if the conditional distribution of the inheritance pattern at the disease locus, given phenotypes, is very different from the prior inheritance probabilities that are deduced from Mendel's law of segregation. See Appendix D for more details. For the type of sibling families considered in our simulations, data sets corresponding to a weak genetic model have larger values of C than data sets corresponding to a strong genetic model (although the strong genetic model has a larger η). We also compare the two score functions S_{pairs} and S_{all}, introduced by Whittemore and Halpern [1994], for a number of sibling families. For the families we consider, the performance of S_{pairs} and S_{all} is quite similar, using $C\eta^2$ as performance criterion, whereas S_{all} does a bit better in terms of η.

SLOPE-TO-NOISE RATIOS

Consider a data set of N pedigrees, where $Z_i(t)$ is the family score of the i^{th} pedigree at locus t, normalized to have zero mean and unit variance under the null hypothesis H_0 that t is unlinked to the disease locus. We assume that the family scores are defined conditionally on observed phenotypes, so that the random variation in $Z_i(t)$ comes from the (perfect) marker data only. In NPL analysis, this usually means that we condition on

the affection status of all pedigree members (i.e., those who have known affection status). However, our setup is valid for arbitrary kinds of phenotypes, both dichotomous (affected/unaffected) and quantitative. We refer to Hössjer [2001a,b] for details.

Following Kruglyak et al. [1996], we define the total linkage score as

$$Z(t) = \sum_{i=1}^{N} \gamma_i Z_i(t),$$

where the weights γ_i satisfy $\sum_1^N \gamma_i^2 = 1$, so that $Z(t)$ has zero mean and unit variance under H_0.

Under the alternative hypothesis H_1 we assume there exists a disease susceptibility locus at some position τ on the chromosomal region(s) of interest. The power to reject H_0, as well as the precision of estimating τ under H_1, depend on the strength of the genetic model, the observed phenotypes, the number of pedigrees, and their graphical structures. Following Feingold et al. [1993], we introduce the noncentrality parameter

$$\eta = \frac{E(Z_N(\tau)|H_1)}{\sqrt{V(Z_N(\tau)|H_0)}} = E(Z_N(\tau)|H_1). \qquad (1)$$

As mentioned in the Introduction, this quantity is related to the power to detect linkage.

Figure 1 gives an example of the observed linkage score function $Z(t)$ and the mean linkage score $E(Z(t))$. Figure 1 displays four simulations of N affected sib pairs (ASPs) with mean family score $\delta = E(Z_i(\tau))$ at the disease locus. The mean sharing score function is used. Hence $Z_i(t)$ equals $-\sqrt{2}$, 0 or $\sqrt{2}$, depending on whether the i^{th} ASP shares 0, 1, or 2 alleles identical by descent at locus t. Since all pedigrees have the same structure, we use equal weights $\gamma_i \equiv N^{-1/2}$. Each row of Figure 1 displays two linkage scores with the same noncentrality parameter $\eta = \sqrt{N}\delta$. There are more crossovers present for the weaker genetic model, corresponding to a smaller δ (and larger N). Notice that the expected linkage score is peaked at the disease locus at 75 cM. This is true also for general types of pedigrees and score functions, and can be formalized as follows. We consider the *local scaling* of the mean score function around τ, and assume there exists a constant $a > 0$ such that

$$E(Z(t) - Z(\tau)) \approx -a|t - \tau| \qquad (2)$$

when t is close to τ. Similarly, the local scaling of the variance function around τ can be described by assuming that

$$V(Z(t) - Z(\tau)) \approx \sigma^2|t - \tau| \qquad (3)$$

for some constant $\sigma^2 > 0$ when t is close to τ. The $\approx$ sign in (2) and (3) means that there are additional terms on the right-hand sides which are of smaller order than $|t - \tau|$ as $t \to \tau$. The ratio of the squared mean slope a^2 and the local variance σ^2 will be

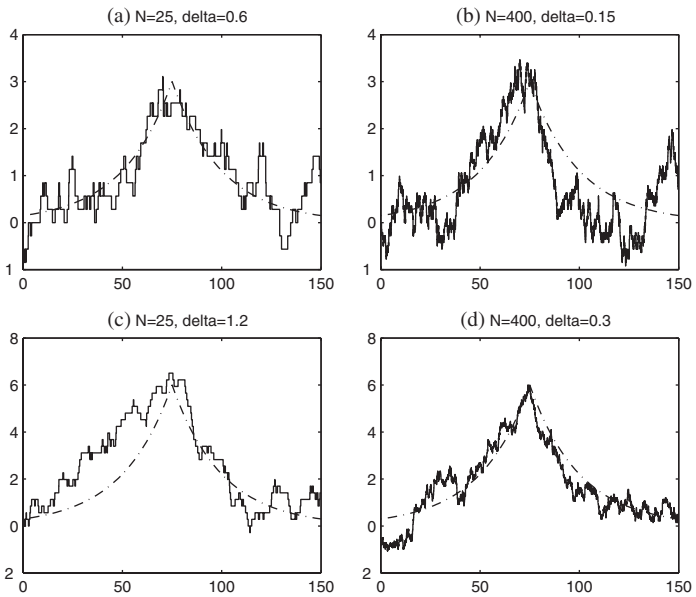

Fig. 1. NPL score Z (solid lines) and $E(Z)$ (dash-dotted lines) under perfect marker information for N affected sib pairs when no interference is assumed for crossovers. Chromosome has length 150 cM, with disease locus τ positioned at 75 cM. Expected family score at τ is δ, and thus noncentrality parameter η equals 3 in a and b, and 6 in c and d.

referred to as the *slope-to-noise ratio* (SLNR) of the linkage score function at τ, and is denoted

$$\text{SLNR} = a^2/\sigma^2. \tag{4}$$

It turns out that SLNR is the right quantity to use for describing the precision with which the disease locus can be mapped. In general, SLNR grows linearly with sample size N. The accuracy with the disease locus can be mapped depends on SLNR^{-1}, and is therefore inversely proportional to the sample size.

To gain further insight about SLNR, we express it in a slightly different way. Let λ be the crossover rate for one meiosis, equal to 1 or 0.01 if map distance is measured in Morgans (M) or cM, respectively. Further, let

$$C = a^2/(\lambda\eta^2\sigma^2)$$

be a *normalized slope-to-noise ratio*, i.e., the ratio of $a/(\lambda\eta)$ squared and σ^2/λ. Here $a/(\lambda\eta)$ is the slope of the mean score function at the disease locus, adjusted for the noncentrality parameter η (average peak height) and crossover rate λ (peak width unit). Similarly, σ^2/λ is the local variance, normalized for λ. We can rewrite SLNR as

$$\text{SLNR} = \lambda C\eta^2 \tag{5}$$

and this formulation shows that SLNR is related to the square of the noncentrality parameter η^2 through the constant C. Since η is an intuitive quantity (average peak height of the linkage score), we believe that (5) gives more insight into the values that SLNR will attain in various situations. It will be seen below that C can vary quite a lot between pedigrees, genetic models, and score functions. This shows that SLNR and η^2 are not equivalent performance criteria. If C is large, the corresponding linkage score is likely to have a sharper peak than another linkage score with a smaller C (but the same η).

Let η_i, a_i, and σ_i^2 be the noncentrality parameter, mean slope, and local variance of the i^{th} family score, defined as (1), (2), and (3), with Z_i in place of Z. Also, let $C_i = a_i^2/(\lambda\eta_i^2\sigma_i^2)$. These quantities are all independent of i when the pedigrees, including their phenotypes, are identical and equal weights $\gamma_i \equiv N^{-1/2}$ are used. This follows from the fact that the family scores Z_i are independent and identically distributed (i.i.d.). In this case, $\eta = \sqrt{N}\delta$, where δ is the common value of all η_i and further, $C = C_1 = \ldots = C_N$ (see Appendix A). Insertion into (5) yields

$$\text{SLNR} = \lambda NC\delta^2. \tag{6}$$

In general, both δ and C will be a function of the pedigree structure, the phenotypes, the genetic model, and the score function.

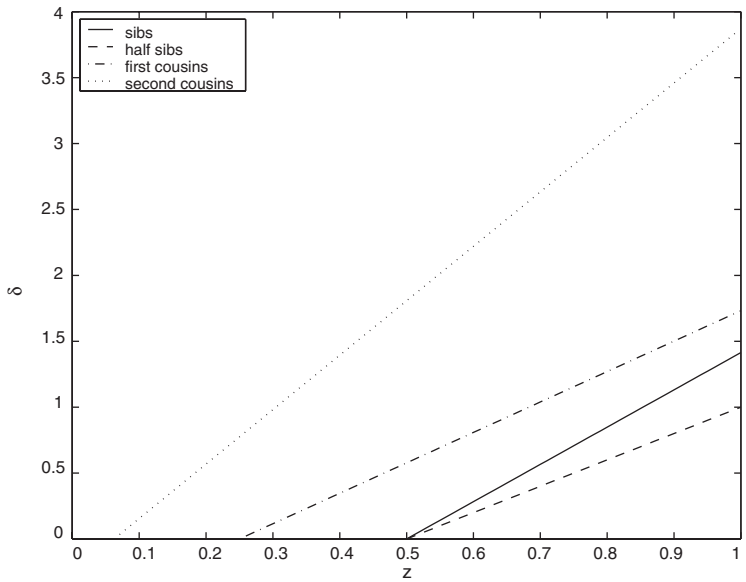

Fig. 2. Plots of noncentrality parameter δ as a function of proportion z of IBD sharing for various affected relative pairs, with score function counting number of alleles shared IBD by pair. Under weak assumptions (nondecreasing penetrances), one has $z_{\min} \le z \le 1$, where $z_{\min}$ is value of z under null hypothesis of no linkage (twice the kinship coefficient). Thus $z_{\min}$ equals 0.5 for sibs, half-sibs, and uncle-nephew pairs, 0.25 for first cousins, and 0.0625 for second cousins. Line for second cousins is identical to that for HBD-sharing of one affected offspring in a first-cousin marriage. In this case, $z_{\min}$ (inbreeding coefficient) is attained for all additive models, whereas $z > z_{\min}$ for recessive models.

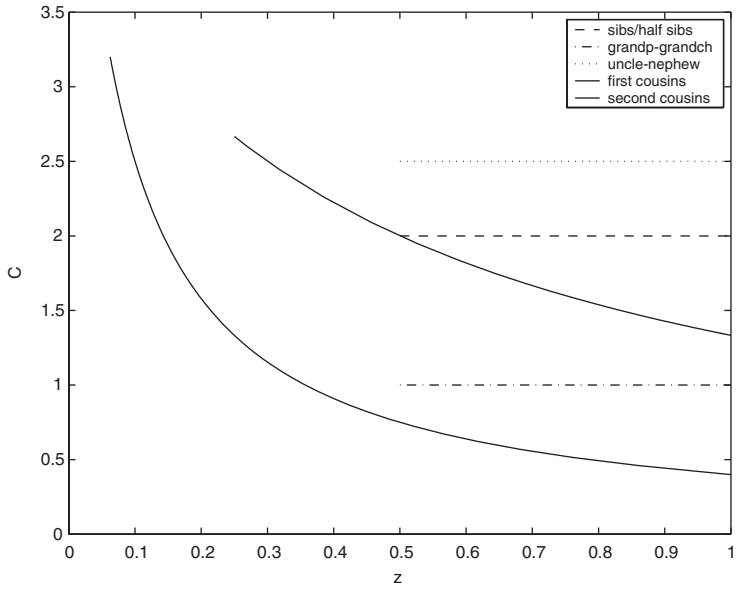

Fig. 3. Plots of *C* as a function of *z* for IBD-sharing of various affected relative pairs. HBD-sharing curve for an offspring of two first cousins is same as IBD-sharing curve for two second cousins.

For dichotomous traits and affected unilineal relative pairs, both δ and C can be written as functions of one parameter, the probability z that the pair shares one allele IBD. The same is true for affected sib pairs when the mean sharing score function is used and z is the fraction of alleles shared IBD by the sibs on average. This can be seen in Figures 2 and 3. Note that C is constant for some pairs, such as sibs, half sibs, uncle-nephew, and grandparent-grandchild, but it varies with z for first and second cousins.

For an inbred pedigree with one affected inbred individual, the score function $S_{\#affHBD}$ [McPeek, 1999] checks whether the inbred individual has its two alleles homozygous by descent (HBD) or not. In this case, both δ and C are functions of z, the probability that the two alleles are IBD; see Figures 2 and 3.

In Figures 4 and 5 values of δ and C are given for several nuclear families with a varying number of affected and unaffected children. One strong dominant and one weak dominant genetic model is considered, along with two score functions, S_{all} and S_{pairs}. It is seen that S_{all} is a bit more powerful in terms of testing (δ) than S_{pairs} for both the strong and weak models. This is compensated by C being larger for S_{pairs}, so that the estimation performance is more similar. In fact, S_{pairs} has slightly better performance than S_{all} in terms of SLNR for the weak genetic model. These differences between the two score functions are interesting and should be further investigated for a larger collection of pedigree structures and genetic models. McPeek [1999], Feingold et al. [2000], and Sengul et al. [2001] compared various score functions in linkage analysis in the testing context in more detail.

CONSTRUCTION OF CONFIDENCE REGIONS

Here, we will describe three confidence regions I_1–I_3 for τ, all having asymptotic coverage probability $P(\tau \in I_j) \to 1-\alpha$ as the number of pedigrees N tends to infinity.

Let $Z_{\max} = \max_t Z(t)$ be the maximal linkage score attained. The *support region* [e.g., Dupuis and Siegmund, 1999] is defined as the set

$$I_1 = \{t;\ Z_{\max} - Z(t) \leq \sigma^2 h_{1\alpha}/a\}.$$

The quantity $\sigma^2 h_{1\alpha}/a$ is (apart from discretization effects) the $(1-\alpha)$-quantile of $Z_{\max} - Z(\tau)$. The factor σ^2/a depends on the pedigree structures, the observed phenotypes, the genetic model, and the score function. The constant $h_{1\alpha}$ is (apart from discretization effects) the $(1-\alpha)$-quantile of a certain random variable defined in Appendix B. When the genetic model is weak, $h_{1\alpha}$ is essentially a function of just α.

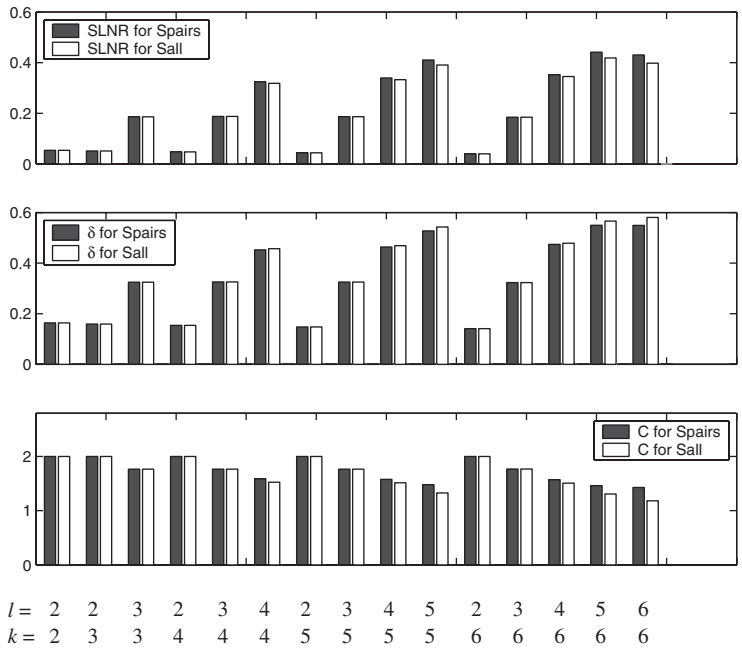

Fig. 4. Values of slope-to-noise ratio SLNR=$\lambda C\delta^2$, noncentrality parameter δ, and constant C for one nuclear family (N=1). Family has k children. Phenotypes are binary, with parents having unknown phenotypes, l children are affected, and $k-l$ unaffected. Crossover rate λ=1. Genetic model is weak dominant, corresponding to a biallelic disease locus with disease allele frequency p=0.0675 and penetrance parameters (f_0, f_1, f_2)=(0.07, 0.3, 0.3). Here f_i is probability of being affected for an individual with i disease alleles. Prevalence is 0.1, and sibling relative risk is 1.29.

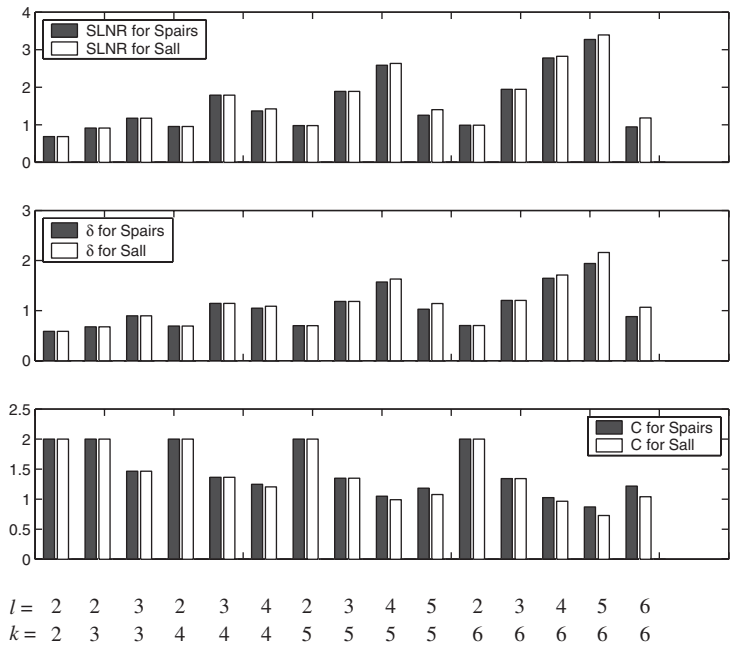

Fig. 5. Same setup as in Figure 4, but for a strong dominant genetic model with p=0.0513, (f_0, f_1, f_2)=(0, 1, 1), prevalence 0.1, and sibling relative risk 5.44.

A disadvantage of the support region is that I_1 need not be an interval. As in Siegmund [1986] and Kruglyak and Lander [1995], we overcome this by introducing the *convex support interval*

$$I_2 = \mathrm{ch}(\{t;\ Z_{\max} - Z(t) \le \sigma^2 h_{2\alpha}/a\})$$

where ch(I) is the convex hull of I, i.e. the smallest interval containing I. To compensate for the convex hull operator, which enlarges the confidence region, $h_{2\alpha}$ is smaller than $h_{1\alpha}$, so that I_2 has asymptotic coverage probability $1-\alpha$ as well. When the genetic model is weak, the quantity $h_{2\alpha}$ essentially depends on α only.

The slope-to-noise ratio is crucial for determining the average length of I_1 and I_2, in the sense that

$$E(|I_j|) \approx \frac{k_{j\alpha}}{\mathrm{SLNR}} = \frac{k_{j\alpha}}{\lambda C \eta^2} \tag{7}$$

for j=1,2, where $|I_j|$ is the length of I_j, and $k_{1\alpha}$ and $k_{2\alpha}$ are constants depending essentially only on α; see Appendix B for definitions. We write $\approx$ in (7), since the right-hand side contains additional terms of smaller order than SLNR^{-1} and η^{-2}, respectively. These remainder terms become negligible in the limit as N grows.

Figure 1 shows that the linkage score is piecewise constant under the perfect marker assumption. This implies that arg max $Z(t)$ is not uniquely defined, but rather a union of finitely many intervals. In order to extract one unique point estimator of τ, we define $\hat{\tau} = (\hat{\tau}^+ + \hat{\tau}^-)/2$, where $\hat{\tau}^+(\hat{\tau}^-)$ is the rightmost (leftmost) point of arg max $Z(t)$. The asymptotic properties of $\hat{\tau}$ as an estimator of τ can be deduced from Kong and Wright [1994] for backcross designs, and from Hössjer [2001a] for general pedigrees, score functions and genetic models. Using $\mathrm{SLNR}(\hat{\tau}-\tau)$ as pivot, this yields an estimation-based confidence interval

$$I_3 = [\hat{\tau} - \mathrm{SLNR}^{-1}k_{3\alpha}/2, \hat{\tau} + \mathrm{SLNR}^{-1}k_{3\alpha}/2]$$

where the constant $k_{3\alpha}$ (defined in Appendix B) is chosen so that I_3 has asymptotic coverage probability $1-\alpha$. As opposed to I_1 and I_2, I_3 has a fixed length. We expect I_3 to be useful mainly for high peaks. Otherwise, I_3 can be located in a valley between two peaks.

Formula (7) has important consequences for planning linkage studies. Let A=SLNR/N be the average slope-to-noise ratio of the sample. In general, A will depend on N (its limit is referred to as the asymptotic slope-to-noise ratio in Hössjer [2001a]), but for i.i.d. family scores, $A = a_i^2/\sigma_i^2 = \lambda C \delta^2$ because of (6). The sample size required to obtain a confidence region I_j with coverage probability $1-\alpha$ and (average) length L is asymptotically

$$N = \frac{k_{j\alpha}}{AL} = \frac{k_{j\alpha}}{\lambda L C \delta^2}, \quad j = 1,2,3 \tag{8}$$

for i.i.d. family scores as $L \to 0$. The relative performance of I_1, I_2, and I_3 is reflected through the constants $k_{1\alpha}$, $k_{2\alpha}$, and $k_{3\alpha}$. Note that the required sample size is asymptotically inversely proportional to the length of the confidence region, the square of the noncentrality parameter δ, and C.

AFFECTED RELATIVE PAIRS

We will now illustrate the behavior of I_1–I_3 for a collection of N identical affected relative pairs which are either sibs or related through just one parent (unilineal relationship). This is essentially the same scenario as considered by Kruglyak and Lander [1995]. Everything herein is also valid for HBD-sharing of a single inbred individual. Let

$$\pi = \frac{\delta}{\delta_{\max}} = \frac{z - z_{\min}}{1 - z_{\min}} \tag{9}$$

be a number which varies between 0 (no linkage) and 1 (IBD- or HBD-sharing at the disease locus can be determined unambiguously from the phenotype). Here $\delta_{\max} = \sqrt{2}$ and $\delta_{\min} = 0.5$ for sibs $\sqrt{3}$ and 0.25 for first cousins, and so on (see Fig. 2).

The quantities $h_{j\alpha}$ and $k_{j\alpha}$, defined above, are functions of the single parameter π for affected relative pairs. In Hössjer [2002], we performed extensive simulations for α=0.5 and 0.05, and found that $h_{1\alpha}$ and $h_{2\alpha}$ are strictly decreasing functions, with the maximum values attained at π=0 and the minimum value zero at π=1. The quantiles $k_{1\alpha}$, $k_{2\alpha}$, and $k_{3\alpha}$ were almost independent of π up to about π=0.5 and then slowly decreased or increased, depending on the value of α and the confidence region method j. The almost constant values of $k_{1\alpha}$, $k_{2\alpha}$, and $k_{3\alpha}$ up to about π=0.5 were around 0.63, 0.94, and 0.79 for α=0.5 and 3, 3.4, and 5.4 for α=0.05. Hence, the average length of the confidence interval is fairly independent of the genetic model as long as it is weak (π is small). We will see below that a Gaussian approximation can be used for weak genetic models. Our simulations thus indicate that this approximation has some robustness.

The behavior of I_1–I_3 has been analyzed by simulation for a collection of N sib pairs with noncentrality parameter δ. The asymptotic approximation of the coverage probability in Table I is quite accurate for I_2 when η is 3 or larger. Slightly larger values of η are needed for I_1, and even larger ones for I_3. In Figure 6, we plotted the expected length of I_2 against η for a 95%

TABLE I. Values of True α, for $(1-\alpha)$ Confidence Regions Based on N Affected Sib Pairs With Noncentrality Parameter δ[a]

			α=0.5			α=0.05			α=0.01		
N	δ	η	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
50	0.1	0.71	0.35	0.24	0.38	0.000	0.000	0.000	0.000	0.000	0.000
100	0.1	1.00	0.49	0.37	0.56	0.008	0.001	0.000	0.000	0.000	0.000
200	0.1	1.41	0.57	0.46	0.62	0.041	0.011	0.063	0.003	0.000	0.000
500	0.1	2.24	0.58	0.52	0.61	0.083	0.043	0.204	0.017	0.005	0.102
1,000	0.1	3.16	0.56	0.52	0.57	0.080	0.056	0.156	0.019	0.011	0.091
50	0.2	1.41	0.57	0.46	0.62	0.038	0.011	0.060	0.002	0.000	0.000
100	0.2	2.00	0.59	0.51	0.62	0.077	0.034	0.197	0.013	0.003	0.070
200	0.2	2.83	0.57	0.52	0.58	0.085	0.056	0.181	0.019	0.010	0.103
500	0.2	4.47	0.52	0.51	0.53	0.064	0.057	0.095	0.014	0.011	0.042
1,000	0.2	6.32	0.51	0.51	0.50	0.057	0.052	0.066	0.013	0.011	0.021
50	0.4	2.83	0.56	0.52	0.58	0.078	0.052	0.178	0.018	0.009	0.096
100	0.4	4.00	0.53	0.51	0.54	0.066	0.056	0.107	0.015	0.012	0.050
200	0.4	5.66	0.51	0.51	0.51	0.056	0.052	0.070	0.012	0.011	0.023
500	0.4	8.94	0.50	0.50	0.50	0.052	0.052	0.056	0.011	0.011	0.015
1,000	0.4	12.65	0.50	0.50	0.49	0.052	0.052	0.053	0.011	0.011	0.013
50	1.0	7.07	0.51	0.50	0.51	0.053	0.052	0.059	0.011	0.010	0.015
100	1.0	10.00	0.50	0.50	0.50	0.050	0.049	0.053	0.011	0.011	0.012
200	1.0	14.1	0.50	0.50	0.50	0.049	0.049	0.049	0.009	0.009	0.010
500	1.0	22.36	0.50	0.50	0.50	0.049	0.050	0.050	0.010	0.010	0.010
1,000	1.0	31.63	0.50	0.50	0.50	0.051	0.050	0.049	0.011	0.011	0.010

[a]Nominal values are α=0.5, 0.05, and 0.01, respectively. Estimates are computed from 50,000 simulated confidence regions when disease locus is positioned in middle of a chromosome of length 150 cM.

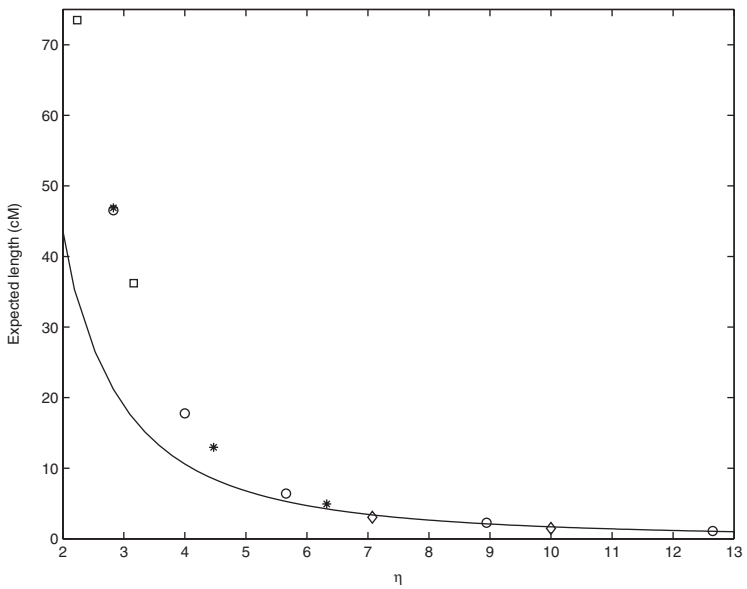

Fig. 6. Solid line is expected length in cM of a 95% convex support region plotted against noncentrality parameter for affected sib pairs, using asymptotic formula (7) with C=2. δ=0.4 is used for computing $k_{2,0.05}$. (Corresponding curves for other δ are virtually indistinguishable.) Also plotted are simulated expected lengths for same combinations of N and δ as in Table I, with δ having values 0.1 (squares), 0.2 (asterisks), 0.4 (open circles) and 1 (diamonds).

confidence region. The asymptotic approximation ignores the curvature of $E(Z)$ in either direction from τ. This is more severe for smaller η, since the confidence region then extends over a larger part of the chromosome. It is seen from Figure 6 that the asymptotic approximation is close for $\eta \geq 4$ and very accurate for $\eta \geq 5.5$. In Table II, we computed the number of sib pairs required to produce 50%, 95%, and 99% confidence regions of expected length 5 cM for various δ. This was done

TABLE II. Required Number of ASPs N in Order for Confidence Regions I_1–I_3 to have (Expected) Length L=5 cM[a]

	α=0.5			α=0.05			α=0.01		
δ	I_1	I_2	I_3	I_1	I_2	I_3	I_1	I_2	I_3
0.05	2,500	3,800	3,000	13,000	14,000	22,000	19,000	21,000	39,000
0.1	630	950	770	3,100	3,500	5,500	4,700	5,100	9,800
0.2	160	240	190	760	850	1,360	1,200	1,200	2,400
0.3	70	100	87	340	380	600	510	550	1,100
0.5	25	36	31	120	140	210	180	200	380
0.7	13	18	16	62	68	100	94	100	180
1.0	7	9	7	29	31	44	43	44	80
$\sqrt{2}$	5	5	4	10	10	15	10	10	23

[a]The (asymptotic) coverage $1-\alpha$ and noncentrality parameter δ for one ASP is varied. Figure 6 (and similar figures for α=0.5 and 0.01 not reported here) reveals that values of N are a bit too small. (N is too small by a factor slightly less than 2 for α=0.5, but is more accurate for α=0.05 and 0.01.) The smaller L is, the more accurate is the asymptotic approximation.

by combining (8) with simulated values of $k_{j\alpha}$ [for details, see Hössjer, 2002].

When the disease locus has not yet been mapped, it might be preferable to express δ (or π) in terms of relative risk ratios, which can be estimated by an epidemiologic study. For instance, Risch [1990] showed that

$$\delta = \frac{\lambda_M - 1}{2\sqrt{2}\lambda_S}$$

for affected sib pairs, where λ_M and λ_S are the risk ratios of a monozygous (MZ) twin and a sibling of an affected individual, respectively. Similarly, δ can be written as a function of relative risk ratios for other affected relative pair [Risch, 1990; Kruglyak and Lander, 1995].

WEAK GENETIC MODELS

For weak genetic models, a large number of pedigrees N is needed to attain a given slope-to-noise ratio. For instance, when the pedigrees have identical structure and phenotypes, this corresponds to letting $\delta \to 0$ and $N \to \infty$, while keeping the relation $\eta = \sqrt{N}\delta$. By a central limit argument, it is then reasonable to approximate Z and its local expansion around τ with Gaussian processes. A discussion of general phenotypes, score functions, and genetic models can be found in Appendix D. See also Feingold et al. [1993].

The Gaussian approximation makes the constants $h_{j\alpha}$ and $k_{j\alpha}$ independent of the genetic model, the score function, and the pedigree. In fact, the quantiles $h_{1\alpha}$, $h_{2\alpha}$, and $k_{3\alpha}$, which are needed to define the three confidence regions, all have explicit expressions (see Appendix D). From simulations [details in Hössjer, 2002], we deduced $k_{2,0.5}$=0.94 and $k_{2,0.05}$=3.47. By plugging these values into (7),we find that the expected lengths of a 50% and 95% convex support region are asymptotically $94/(C\eta^2)$ cM and $347/(C\eta^2)$ cM, respectively, as η grows.

NONIDEAL CONFIDENCE REGIONS

The confidence regions I_1–I_3 are ideal, in the sense that SLNR $h_{1\alpha}$, $h_{2\alpha}$, and $k_{3\alpha}$ depend on the genetic model. Therefore, they should mainly be used for planning linkage studies. Given a certain genetic model and score function, how many pedigrees are required to obtain a region of given length and confidence level?

It would be interesting to use plug-in versions of I_1–I_3, with the genetic model estimated from data. For ASPs, this entails estimating η, since all unknown quantities depend on η (or equivalently δ). Since η is the maximal value of $E(Z(t))$, it would be tempting to use $Z_{\max}$ as an estimate of η. However, this will produce an upward bias. Similar phenomena were recently discussed for QTL models by Göring et al. [2001] and Allison et al. [2002]. Interestingly, the asymptotic expansions in Appendix B can be used to construct at least a first-order bias correction of $Z_{\max}$ as an estimate of η. The basic expansion to use is

$$E(Z_{\max}) \approx g(\eta) \tag{10}$$

where $g(\eta) = \eta + 2M(\eta/\sqrt{2N})/\eta$. This approximation is accurate when η becomes large; see Appendix B for derivation as well as a definition of function M. An asymptotic method-of-moments estimate of η is thus

$$\hat{\eta} = g^{-1}(Z_{\max}).$$

This bias correction does not take into account the size of the chromosomal region where linkage

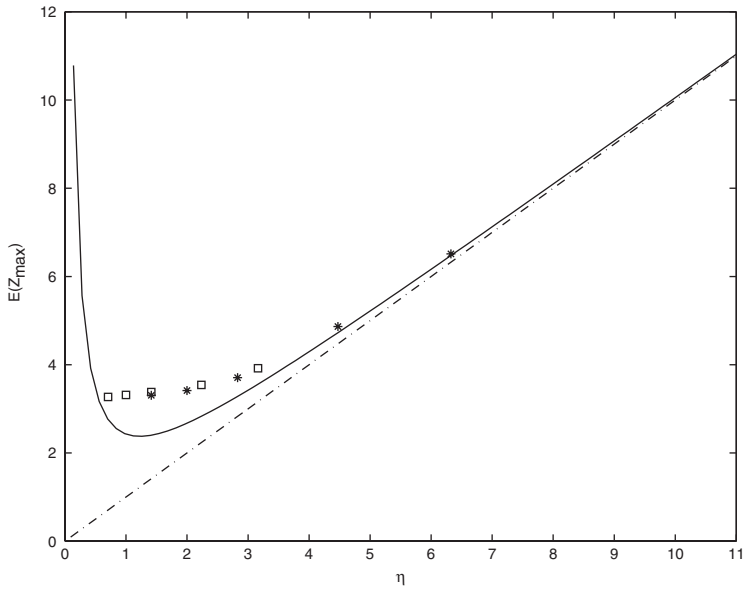

Fig. 7. Expected maximal linkage score as a function of noncentrality parameter η for affected sib pairs. Solid curve is approximation $g(\eta)$ in (10) with N=100 (corresponding curves for other N are virtually indistinguishable). Also shown are estimates of genomewise expected maximal linkage scores. Each estimate is based on 1,000 simulated genomewise linkage scores, with sex averaged chromosome lengths taken from Table 1 in Ott [1999]. Disease locus is positioned at midpoint of chromosome 1 (which has length of 298.5 cM). Same combinations of δ and N as in Table I are used, with δ=0.1 (squares), δ=0.2 (asterisks), δ=0.4 (open circles), and δ=1.0 (diamonds).

is sought. Instead, it is solely based on the random fluctuation of the linkage score locally around the disease locus. Figure 7 displays the asymptotic approximation (10) as well as simulated genomewide estimates of $E(Z_{\max})$. The asymptotic approximation is quite accurate for genomewide scans when the noncentrality parameter η is 4 or larger.

DISCUSSION

In this paper, we have defined three kinds of confidence regions. Their construction depends on the underlying genetic model and they are in that sense ideal, although we briefly discussed nonideal plug-in versions to use when the genetic model is unknown. The main application is planning of linkage studies. An asymptotic approximation of the expected length of a confidence region with a given level of confidence can be computed under a perfect marker assumption. The answer depends on the genetic model, the score function, the pedigree structures, and the weighting scheme of the pedigrees.

The definitions of all three confidence regions involve the quantiles, $h_{1\alpha}$, $h_{2\alpha}$, and $k_{3\alpha}$ of certain distributions defined in Appendix B. We have discussed how to compute these quantiles explicitly for affected relative pair families. For common designs, with pedigrees of different structures, the quantiles have quite complicated expressions. In principle, they can be computed straightforwadly by simulation, although this can be time-consuming. On the other hand, our results for affected relative pairs indicate that the quantiles are very robust towards changes of the genetic model parameters for weak or moderately weak genetic models. In addition, many linkage studies for complex diseases involve a contribution from several genes, with each gene having a small individual effect. Under such circumstances, it is reasonable to use the weak genetic model approximations based on Gaussian processes. Then explicit expressions for the quantiles of interest are available (independent of the pedigree structures, genetic model, score function, and weighting scheme), as discussed in Appendix D.

The perfect marker assumption has been used throughout this paper. In practice, this assumption is violated because 1) genetic markers have a finite spacing with heterozygosity lower than one, 2) not all family members are genotyped, and 3) several founders may be related and share alleles identical by descent. In view of the large number of

genetic markers currently available, assumption 1 is less problematic. This is because linkage analysis is based on crossovers occuring in the pedigrees (not in previous generations). For instance, a medium-sized pedigree with 20 meioses has on average 1 crossover per 5 cM. Therefore, a marker map with interdistance 1 cM will have a good chance of detecting all or most crossovers provided that the markers are reasonably polymorphic. A marker map with interdistance 1 cM can be attained, for instance, if linkage analysis is performed in two steps. An initial genome scan with a coarse grid of markers is followed by a second scan from interesting regions pinpointed by the first scan. In the second scan, which is defined for a much smaller region, a finer marker map is employed. Assumption 2 is more problematic, especially for large extended pedigrees. The perfect marker assumption can be used if the value of the score function does not change because of missing phenotype and marker information from untyped pedigree members. The simplest example is the mean sharing score function for an affected sib pair. This score function only requires genotyping of the sib pair, and not of the parents, for its definition. (However, if only the sibs are genotyped, a denser map of markers is needed for the perfect marker assumption to be accurate.) Assumption 3 can be problematic in families from isolated regions when some founders have a recent ancestor.

In any case, the perfect marker approach gives a lower bound for the true expected length of the confidence regions. The true expected length will depend on the positioning of the genetic markers in close vicinity of the disease locus, as well as their heterozygosity. As a rule of thumb, the interdistance between markers should be of a lower order than the length of the confidence regions, in order for the perfect marker approximation to be accurate.

The results in this paper are based on asymptotic approximations. Our simulations show that these approximations are accurate when the confidence regions have a length of about 5 cM or shorter, or when the noncentrality parameter is about 5 or larger. Mathematically, we look at the linkage score locally around the disease locus, and the asymptotic approximations require that the local neighborhood be small compared to the whole chromosome. In fact, the asymptotic approximations are based on two local quantitities computed at the disease locus, i.e., the slope of the mean linkage score a, and the amount of random fluctuations σ^2. The upward curvature of the mean linkage score is ignored. This results in a too-optimistic confidence region, since the mean linkage score is underestimated at regions far off from the disease locus.

We briefly discussed nonideal plug-in versions of our confidence regions for affected sib pairs. This can be done more generally for other pedigree structures, phenotypes, and genetic models. For each particular genetic model, one identifies the relevant parameters θ and investigates how the expected maximum linkage score $E(Z_{\max}) = g(\theta)$ depends on these parameters. Then $\hat{\theta} = g^{-1}(Z_{\max})$ gives a first-order bias-corrected estimate of θ. When the genetic model is weak, we can rely on the Gaussian approximation in Appendix D. This makes computation of g more feasible. Siegmund [2002] used the Gaussian approximation for correcting for upward bias. His approach differs from ours in that he uses quantiles of $Z_{\max}$ as a function of θ rather than the expected value of the maximum linkage score.

We assumed a single-locus genetic model. Everything treated in this paper can be generalized to multilocus models if the other disease loci are unlinked to τ. The crucial part is to generalize the probability $P_i(w)$ defined in Appendix A to multilocus models; see Hössjer [2000a] for details.

ACKNOWLEDGMENTS

The author thanks the editor and two referees for many helpful suggestions which improved the presentation of this paper.

APPENDIX A

COMPUTATION OF SLOPE-TO-NOISE RATIOS AND C

Let a_i and σ_i^2 be defined as before (6). Then, by (4) and the linearity of the mean and variance functions

$$\mathrm{SLNR} = \frac{\left(\sum_{i=1}^N \gamma_i a_i\right)^2}{\sum_{i=1}^N \gamma_i^2 a_i^2}. \qquad (11)$$

Analogously, the noncentrality parameter can be expressed as $\eta = \sum_{i=1}^N \gamma_i \eta_i$. This in turn gives $C = \mathrm{SLNR}/(\lambda\eta^2)$.

For equal weights $\gamma_i \equiv 1/\sqrt{N}$ and i.i.d. family scores, (11) reduces to (6). Further, $\eta = \sqrt{N}\eta_i$ and $C_i = C$.

The quantities a_i and σ_i^2 can be computed as follows: let $v_i(t)$ be the inheritance vector of the i^{th} pedigree at locus t. If there are n_i members and f_i founders in this pedigree, $v_i(t)$ is a binary vector of length twice the number of nonfounders $m_i = 2(n_i - f_i)$. Each bit is 0 (1) if during the corresponding meiosis a grandpaternal (grandmaternal) allele is transmitted [Kruglyak et al., 1996]. We assume that the score function $S_i : \{0,1\}^{m_i} \to \mathbb{R}$ for the i^{th} pedigree has been normalized to have zero mean and unit variance under H_0, i.e., when $v_i(t)$ is uniformly distributed over all 2^{m_i} possible inheritance vectors. Thus $Z_i(t) = S_i(v_i(t))$ under perfect marker information. (The inheritance vector $v_i(t)$ is only known up to uncertainty of the phase of all founders [Kruglyak et al., 1996]. However, if the score function S_i is invariant with respect to this uncertainty, $S_i(v_i(t))$ will be known, although $v_i(t)$ is not.) Let $P_i(w) = P(v_i(\tau) = w | Y_i, H_1)$ be the probability function of the inheritance vector at the disease locus *conditional on the observed phenotype vector* Y_i for the i^{th} pedigree. Details on computation of $P_i(w)$ can be found in Kruglyak et al. [1996, Appendix C]. It is shown in Hössjer [2001a] that

$$a_i = \lambda \cdot \left(m_i \sum_w S_i(w) P_i(w) - \sum_{j=1}^{m_i} \sum_w S_i(w + e_j) P_i(w) \right),$$

$$\sigma_i^2 = \lambda \cdot \sum_w P_i(w) \sum_{j=1}^{m_i} \left(S_i(w + e_j) - S_i(w) \right)^2$$

where e_j is the binary vector of length m_i, with one in position j and zeros elsewhere, and + refers to componentwise modulo 2 addition. The corresponding formula for the i^{th} noncentrality parameter is $\eta_i = \sum_w S_i(w) P_i(w)$.

APPENDIX B

LOCAL SCALING OF LINKAGE PROCESS AND CONFIDENCE INTERVALS

We start by rescaling the linkage score as

$$\tilde{Z}_N(s) = a(Z(\tau + s/\text{SLNR}) - Z(\tau))/\sigma^2. \tag{12}$$

Thus we look at the linkage score on a local chromosomal scale $O(\text{SLNR}^{-1}) = O(N^{-1})$ and vertical scale $O(\sigma^2/a) = O(N^{-1/2})$. For large N, the rescaled linkage process $\tilde{Z}_N$ can be well-approximated by a limiting compound Poisson process $\tilde{Z}$; see Hössjer [2001a] for a formal proof. We define $\tilde{Z}$ as follows: $\tilde{Z}(0) = 0$, and then $\tilde{Z}$ is piecewise constant, with jumps occuring in both directions from the origin, according to two independent Poisson processes with intensity $\tilde{\lambda} = \lambda \cdot \text{SLNR}^{-1} \sum_1^N m_i$, where λ is the crossover rate for one meiosis. A jump is defined as the change of $\tilde{Z}$ when we move away *from* the origin, i.e., to the right (left) when $s > 0$ ($s < 0$). All jumps of $\tilde{Z}$ are independent and identically distributed (i.i.d.) with the same distribution as a random variable X, which can be written as

$$X = \frac{a}{\sigma^2} \gamma_J (S_J(v') - S_J(v)) \tag{13}$$

where (J, v, v') is random: $P(J = i) = m_i / \sum_{j=1}^n m_j$, $P(v = w | J = i) = P_i(w)$, and $P(v' = w' | J = i, v = w)$ equals $1/m_i$, if w' and w differ by one bit and are zero otherwise. After some manipulations, using the definitions of a and σ^2, it can be verified that

$$E(\tilde{Z}(s)) = -|s|, \quad V(\tilde{Z}(s)) = |s|. \tag{14}$$

Thus $\tilde{Z}$ drifts away to $-\infty$ in both directions from the origin.

The weak convergence $(\xrightarrow{\mathscr{L}})$ of $\tilde{Z}_N$ towards $\tilde{Z}$ makes it possible to find asymptotically valid expressions for $h_{1\alpha}$, $h_{2\alpha}$, $k_{1\alpha}$, $k_{2\alpha}$, and $k_{3\alpha}$. Note first that $\tau \in I_1$ iff $Z_{\max} - Z(\tau) \le \sigma^2 h_{1\alpha}/a$. This is asymptotically equivalent to $\tilde{Z}_{\max} \le h_{1\alpha}$, where $\tilde{Z}_{\max} = \max_s \tilde{Z}(s)$. Thus we take

$$h_{1\alpha} = (1 - \alpha)\text{-quantile of } \tilde{Z}_{\max}. \tag{15}$$

Note further that $\tau \in I_2$ iff $|\max_{t \ge \tau} Z(t) - \max_{t \le \tau} Z(t)| \le \sigma^2 h_{2\alpha}/a$. This is asympotically equivalent to $|\tilde{Z}_{\max}^+ - \tilde{Z}_{\max}^-| \le h_{2\alpha}$, where $\tilde{Z}_{\max}^+ - \max_{s \ge 0} \tilde{Z}(s)$ and $\tilde{Z}_{\max}^- = \max_{s \le 0} \tilde{Z}(s)$. Thus we pick $h_{2\alpha}$ according to

$$h_{2\alpha} = (1 - \alpha)\text{-quantile of } |\tilde{Z}_{\max}^+ - \tilde{Z}_{\max}^-|. \tag{16}$$

Note that $\tilde{Z}_{\max} = \max(\tilde{Z}_{\max}^+, \tilde{Z}_{\max}^-)$. Further, due to the construction of $\tilde{Z}$, $\tilde{Z}_{\max}^+$ and $\tilde{Z}_{\max}^-$ are independent random variables with the same distribution.

In order to give expressions for $k_{1\alpha}$ and $k_{2\alpha}$, we first define rescaled confidence regions $\tilde{I}_1 = \{s;\ \tilde{Z}_{\max} - \tilde{Z}(s) \le h_{1\alpha}\}$ and $\tilde{I}_2 = \text{ch}(\{s;\ \tilde{Z}_{\max} - \tilde{Z}(s) \le h_{2\alpha}\})$, corresponding to I_1 and I_2. Then set

$$k_{j\alpha} = E(\tilde{I}_j), \quad j = 1, 2. \tag{17}$$

Finally, in order to define $k_{3\alpha}$, we let $\hat{s}^-$ and $\hat{s}^+$ be the left- and rightmost maximum points of $\tilde{Z}$ and $\hat{s}(\hat{s}^+ + \hat{s}^-)/2$. The weak convergence $\tilde{Z}_N \xrightarrow{\mathscr{L}} \tilde{Z}$ implies that $\text{SLNR}(\hat{\tau} - \tau) \xrightarrow{\mathscr{L}} \hat{s}$, see Hössjer, [2001a] for details. This gives an asymptotically valid distribution for the pivot $\text{SLNR}(\hat{\tau} - \tau)$, and

$$k_{3\alpha} = 2 \cdot ((1 - \alpha/2)\text{-quantile of } \hat{s}). \tag{18}$$

APPENDIX C

AFFECTED RELATIVE PAIRS

As a special case of the general framework in Appendices A and B, consider a collection of N affected relative pairs, with $0<\pi\leq 1$ as defined in (9). The limiting process $\tilde{Z}$ is then a Poisson-inbedded two-sided random walk with negative drift, whose distribution only depends on π. This class of processes was previously considered by Kong and Wright [1994] for backcrosses, and by Kruglyak and Lander [1995] for affected relative pairs.

The intensity of the jumps equals $\tilde{\lambda}=1/\pi^2$, and the jumps in (13) have a two-point distribution with $P(X=-\pi)=(1+\pi)/2$, $P(X=\pi)=(1-\pi)/2$. Explicit expressions for $h_{1\alpha}$ and $h_{2\alpha}$ in (15) and (16) are available, by noting that the maximum of $\tilde{Z}/\pi$ in both directions from the origin has the same distribution as the maximum of a random walk starting at the origin and having probability $(1-\pi)/2$ and $(1+\pi)/2$ of upward (+1) and downward (−1) jumps, respectively. It can be shown [see Appendix D in Kruglyak and Lander, 1995] that this implies that $\tilde{Z}^+_{\max}/\pi$ (and $\tilde{Z}^-_{\max}/\pi$) have geometric distributions, i.e., $P(\tilde{Z}^+_{\max}=\pi i)=(1-p)^i p$, $p=2\pi/(1+\pi)$, and $i=0,1,2,\ldots$ After some algebra, this implies

$$P(\tilde{Z}_{\max}\leq h)=(1-(1-p)^{[h/\pi]+1})^2 \qquad (19)$$

and

$$P(|\tilde{Z}^+_{\max}-\tilde{Z}^-_{\max}|\leq h)=\begin{cases}\pi, & 0\leq h<\pi,\\ 1-\left(\dfrac{1-\pi}{1+\pi}\right)^{[h/\pi]+1}(1+\pi), & h\geq\pi.\end{cases} \qquad (20)$$

where $[x]$ is the largest integer smaller than or equal to x. In principle, $h_{1\alpha}$ and $h_{2\alpha}$ can be computed as the $(1-\alpha)$-quantiles of the distribution functions in (19) and (20), respectively. Note, however, that $\tilde{Z}_{\max}$ and $|\tilde{Z}^+_{\max}-\tilde{Z}^-_{\max}|$ have discrete distributions. This will create a bias when computing $k_{1\alpha}$ and $k_{2\alpha}$ according to (17). Especially for large δ, this effect is quite pronounced. To avoid this, we will actually compute the length $|\tilde{I}_1|$ (the reasoning for $\tilde{I}_2$ is analogous) as follows. Given α, define numbers h_{11} and h_{12} such that $p_1=P(\tilde{Z}_{\max}\leq h_{11})<1-\alpha\leq p_2=P(\tilde{Z}_{\max}\leq h_{12})$. Let $\tilde{I}_{11}$ and $\tilde{I}_{12}$ be the confidence regions obtained when $h_{1\alpha}$ is replaced by h_{11} and h_{12}, respectively, in the definition of $\tilde{I}_1$. Then put

$$|\tilde{I}_1|=(1-r)|\tilde{I}_{11}|+r|\tilde{I}_{12}| \qquad (21)$$

where $r=(1-\alpha-p_1)/(p_2-p_1)$. This corresponds to using a randomized decision rule when constructing the confidence region, so that the asymptotic confidence level is exactly α. Then $k_{1\alpha}$ is computed from (17) by repeatedly simulating confidence regions of "length" (21). We also use $h_{1\alpha}=(1-r)h_{11}+rh_{12}$ when plotting $h_{1\alpha}$ as a function of π.

The constant $k_{3\alpha}$ will be determined using (18) and simulation. When $\pi=1$, an analytical solution is easily available, since $\hat{s}^+$ and $\hat{s}^-$ are independent exponential random variables with mean $\pi^2=1$. From this, it follows that $k_{3\alpha}=-\ln(\alpha)$ when $\pi=1$.

We end this appendix by motivating the expansion (10) for affected sib pairs. The distribution function of $\tilde{Z}_{\max}$ is given in (19) as π times the maximum of two independent geometric distributions. After some computations, one obtains

$$M(\pi):=E(\tilde{Z}_{\max})=\tfrac{1}{4}(1-\pi)(3+\pi)$$

which is a decreasing function of π, with $M(0)=3/4$ and $M(1)=0$. The rescaling (12) and the fact that $\tilde{Z}_N$ is close to $\tilde{Z}$ in distribution implies

$$\begin{aligned}E(Z_{\max})&=E(Z(\tau))+\tfrac{a}{\sigma^2}E(\tilde{Z}_{\max})+o(N^{-1/2})\\&=\eta+\tfrac{2}{\eta}M(\eta/\sqrt{2N})+o(\eta^{-1}).\end{aligned}$$

Here we used the fact that $\eta=\sqrt{N}\delta=\sqrt{2N}\pi$, $a=\sqrt{N}a_i=4\sqrt{N}\lambda\delta$, $\sigma^2=\sigma_i^2=8\lambda$, and $a/\sigma^2=\eta/2$ for affected sib pairs. By combining the last two displayed equations, we obtain an explicit expression for g in (10).

APPENDIX D

WEAK GENETIC MODELS

A weak genetic model corresponds to the case when the conditional inheritance distribution at the disease locus is close to uniform, i.e., $P_i(w)=2^{-m_i}(1+\varepsilon R(w)+o(\varepsilon))$, where $R(w)$ is the likelihood score function, and ε is a small positive number [see Whittemore, 1996; Hössjer, 2001b]. When $\varepsilon\to 0$, the process $\tilde{Z}$ has jumps X of size tending to zero, and intensity $\tilde{\lambda}$ tending to infinity in such a way that (14) is maintained. By a central limit theorem argument, it follows that in the limit we obtain the Gaussian process

$$\tilde{Z}(s)=B(s)-|s|$$

where B is a two-sided standard Brownian motion. For affected relative pairs, this corresponds to the limit $\pi\to 0$.

The Gaussian framework simplifies some formulas. To start with, $\tilde{Z}^+_{\max}$ and $\tilde{Z}^-_{\max}$ are

independent exponential random variables with mean 0.5. From this, it follows that $h_{1\alpha} = -\ln(1-\sqrt{1-\alpha})/2$, and $h_{2\alpha} = -\ln(1-\alpha)/2$. Further, an explicit formula for $k_{3\alpha}$ can be obtained from Siegmund [1986].

REFERENCES

Allison DB, Fernandez JR, Heo M, Zhu S, Etzel C, Beasly TM, Amos C. 2002. Bias in estimates of quantitative-trait-locus effect in genome scan: Demonstration of the phenomenon and a method-of-moments procedure for reducings bias. Am J Hum Genet 70:575–85.

Clerget-Darpoux F, Bonaïti-Pellié C, Hochez J. 1986. Effects of misspecifying genetic parameters in lod score analysis. Biometrics 42:393–9.

Darvasi A, Soller M. 1997. A simple method to calculate resolving power and confidence interval of QTL map location. Behav Genet 27:125–32.

Darvasi A, Weinreb A, Minke V, Weller JI, Soller M. 1993. Detecting marker-QTL linkage and estimating QTL gene effect and map position using a saturated genetic map. Genetics 134:943–51.

Dupuis J, Siegmund D. 1999. Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373–86.

Feingold E, Brown PO, Siegmund D. 1993. Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent. Am J Hum Genet 53:234–51.

Feingold E, Song KK, Weeks DE. 2000. Comparison of allele-sharing statistics for general pedigrees. Genet Epidemiol [Suppl] 19:92–8.

Göring H, Terwilliger J, Blangero J. 2001. Large upward bias in estimation of locus-specific effects from genomewide scans. Am J Hum Genet 69:1357–69.

Hössjer O. 2001a. Preprint 2001:16, Centre for Mathematical Sciences, Lund University, Sweden. Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. Ann Stat. (in press).

Hössjer O. 2001b. Preprint 2001:17, Centre for Mathematical Sciences, Lund University, Sweden. Determining inheritance distributions via stochastic penetrances. J Am Stat Assoc. (in press).

Hössjer O. 2002. Assessing accuracy in linkage analysis by means of confidence regions. Preprint 2002:17, Centre for Mathematical Sciences, Lund University, Sweden.

Kong A, Wright F. 1994. Asymptotic theory for gene mapping. Proc Natl Acad Sci USA 91:9705– 9.

Kruglyak L. 1997. The use of a genetic map of biallelic markers in linkage studies. Nat Genet 17:21–4.

Kruglyak L, Daly MJ. 1998. Linkage thresholds for two-stage genome scans. Am J Hum Genet 62:994–6.

Kruglyak L, Lander ES. 1995. High-resolution gene mapping of complex traits. Am J Hum Genet 56:1212–23.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–63.

Lander EL, Kryglyak L. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–7.

Liang KY, Chiu Y-F, Beaty TH. 2001. A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. Hum Hered 51:64–78.

McPeek MS. 1999. Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 16:225–49.

Nilsson S. 1999. Two contributions to genetic linkage analysis. Licentiate thesis, Chalmers University of Technology, Gothenburg, Sweden.

Ott J. 1999. Analysis of human genetic linkage, 3rd ed. Baltimore: Johns Hopkins Univiversity Press.

Risch N. 1990. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–8.

Sengul H, Weeks DE, Feingold E. 2001. A survey of affected-sibship statistics for nonparametric linkage analysis. Am J Hum Genet 69:179–90.

Siegmund D. 1986. Boundary crossing probabilities and statistical applications. Ann Stat 14:361–404.

Siegmund D. 2002. Upward bias in estimation of genetic effects. Am J Hum Genet 71:1183–8.

Terwillinger JD, Ding Y, Ott J. 1992. On the relative importance of marker heterozygosity and intermarker distance in gene mapping. Genomics 13:951–6.

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

Weeks DE, Lange K. 1992. A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859–68.

Whittemore AS. 1996. Genome scanning for linkage: an overview. Biometrics 50:118–27.

Whittemore AS, Halpern J. 1994. A class of tests for linkage using affected pedigree members. Biometrics 50:118–27.