Biostatistics (2005), **6**, 4, *pp.* 520–538
doi:10.1093/biostatistics/kxi025
Advance Access publication on April 14, 2005

Improving the calculation of statistical significance in genome-wide scans

LARS ÄNGQUIST*

Department of Mathematical Statistics, Lund University, Lund, Sweden
larsa@maths.lth.se

OLA HÖSSJER

Department of Mathematics, Stockholm University, Stockholm, Sweden

SUMMARY

Calculations of the significance of results from linkage analysis can be performed by simulation or by theoretical approximation, with or without the assumption of perfect marker information. Here we concentrate on theoretical approximation. Our starting point is the asymptotic approximation formula presented by Lander and Kruglyak (1995, *Nature Genetics*, **11**, 241–247), incorporating the effect of finite marker spacing as suggested by Feingold *et al.* (1993, *American Journal of Human Genetics*, **53**, 234–251). We consider two distinct ways in which this formula can be improved. Firstly, we present a formula for calculating the crossover rate ρ for a pedigree of general structure. For a pedigree set, these values may then be weighted into an overall crossover rate which can be used as input to the original approximation formula. Secondly, the unadjusted p-value formula is based on the assumption of a Normally distributed nonparametric linkage (NPL) score. This leads to conservative or anticonservative p-values of varying magnitude depending on the pedigree set structure. We adjust for non-Normality by calculating the marginal distribution of the NPL score under the null hypothesis of no linkage with an arbitrarily small error. The NPL score is then transformed to have a marginal standard Normal distribution and the transformed maximal NPL score, together with a slightly corrected value of the overall crossover rate, is inserted into the original formula in order to calculate the p-value. We use pedigrees of seven different structures to compare the performance of our suggested approximation formula to the original approximation formula, with and without skewness correction, and to results found by simulation. We also apply the suggested formula to two real pedigree set structure examples. Our method generally seems to provide improved behavior, especially for pedigree sets which show clear departure from Normality, in relation to the competing approximations.

Keywords: Adjusted approximation formula; Allele sharing; Approximation of distributions; Crossover rate; Deviation from Normality; Extreme value formulas; Genome-wide significance; Hermite polynomials; Marker density; Nonparametric linkage analysis.

*To whom correspondence should be addressed.

1. Introduction

The techniques of linkage analysis are usually used to identify possible chromosomal regions that are linked to a given phenotype (disease). This article will deal with nonparametric linkage (NPL) analysis (cf. e.g. Kruglyak and Lander, 1995; Kruglyak *et al.*, 1996) and the issue of calculating the statistical significance (or p-value) of the maximal NPL score when performing a genome-wide scan. For an overview of different approaches to gene mapping, see Ott (1999).

When calculating appropriate p-values, it is possible to use two distinctly different approaches—simulation or theoretical approximation. In this work we use the latter approach. Our aim is to improve the performance of the asymptotic approximation formula given in Lander and Kruglyak (1995), with the adjustment of finite marker spacing as suggested by Feingold *et al.* (1993). This formula is based on extreme value theory and Normal approximation of stochastic processes, a method originally applied to p-value calculation in linkage analysis by Lander and Botstein (1989). We suggest two different improvements. Firstly, we give a formula for calculating the crossover rate ρ for a general collection of pedigrees based on the theory presented in Hössjer (2003a). Secondly, we approximate the marginal distribution F of the NPL score by F_ϵ, with arbitrarily good accuracy, and transform F_ϵ to the standard Normal distribution function Φ to improve performance of the p-value approximation.

Methods of p-value calculation based on permutation tests are described for quantitative traits and animal breeding models by Doerge and Churchill (1996) and Abney *et al.* (2002).

2. Basic linkage analysis

2.1 *Definitions*

A *pedigree* is a set of relatives. The members of a pedigree may be divided into two subgroups—*founders* and *nonfounders*. The distinction between these groups is based on the fact that the parents of the founders are not included in the pedigree. An individual's *genotype* at a specific marker locus consists of two *alleles*. One allele is inherited from the mother (maternal allele) and the second from the father (paternal allele). The process leading to the formation of gametes (egg or sperm cells), i.e. the mixing of grandpaternal and grandmaternal DNA segments, is called *meiosis*. Further, n is the number of individuals in the pedigree, f the number of founders, $n - f$ the number of nonfounders, and $m = 2(n - f)$ the number of meioses.

The alleles of the founders may be described as $g = (g_1, g_2, \ldots, g_{2f})$ and inheritance in the pedigree may then be seen as the distribution of the founder alleles to the nonfounders. Two individuals are said to share an allele *IBD* (identical-by-descent) if they have inherited exactly the same founder allele ($g_i \in g$). When performing linkage analysis, one is ordinarily using information from a whole *pedigree set* consisting of N distinct pedigrees.

At each marker locus, x, the whole inheritance process for a given pedigree may be described by the *inheritance vector*, $v(x) = (p_1, m_1, p_2, m_2, \ldots, p_{n-f}, m_{n-f})$, where p_i and m_i equal 0 if, at position x, the ith nonfounder's paternal and maternal alleles, respectively, originate from a grandfather and 1 if they originate from a grandmother.

In *NPL analysis*, one defines *genetic linkage* to a locus as deviation from random inheritance of the founder alleles among the subset of pedigree members with a certain phenotype (e.g. the affecteds). To discover possible linkage to a locus within a specific chromosomal region, one tests for linkage against a collection of markers covering the region of interest.

Usually there is incomplete *information* in the data set. Data may be lost, only a subset of the pedigree members may be genotyped and the markers may be nonpolymorphic with a large map distance between adjacent loci. Measures of the information contained in the data set at a locus, usually ranging from 0 (no information) to 1 (full information), include the widely used entropy-based information measure

introduced in Kruglyak *et al.* (1996), a variance-based measure (cf. e.g. Kruglyak and Lander, 1995) which may be thought of as an estimate of the variance of the NPL score, and measures based on Fisher information or LOD scores (log-odds statistic). Recommended references are Nicolae *et al.* (1998) and Nicolae (1999).

A *single-point* linkage analysis uses only the information contained in the data set at each marker locus separately. The natural extension, where one uses information from several neighboring markers (on the same chromosome), is called *multipoint* linkage analysis (Kruglyak and Lander, 1995). Available software packages for analyses of this kind include GENEHUNTER (Kruglyak *et al.*, 1996), ALLEGRO (Gudbjartsson *et al.*, 2000), or MERLIN (Abecasis *et al.*, 2002).

2.2 *Score functions and the NPL score*

To show evidence of linkage at a marker locus, we introduce a *score function*. In general, for a given pedigree, a score function S is a function of the inheritance vector v that quantifies compatibility between v and the phenotypes of the pedigree members. However, in many cases it can also be interpreted in the usual statistical sense as the derivative, with respect to genetic model parameters, of the log-likelihood of marker data given phenotypes (Whittemore, 1996; McPeek, 1999; Hössjer, 2003b, 2005).

If the genetic model is known, it may be possible to find a function that is optimal in terms of power to detect linkage. For example, when considering complex diseases with unknown mode of inheritance, it may be preferable to choose among score functions that are robust with respect to variations in the genetic model. Throughout this work we use the well-known S_{all} function (Whittemore and Halpern, 1994), defined for binary phenotypes as the degree of *IBD* sharing among the affected pedigree members. The literature discussing and evaluating the performance of score functions applied to different models of inheritance is rich and the performance of S_{all} is usually robust, especially when the information content of the data is reasonably high (Whittemore and Halpern, 1994; Kruglyak *et al.*, 1996; Davis and Weeks, 1997; McPeek, 1999; Feingold *et al.*, 2000; Sengul *et al.*, 2001).

From now on we assume that the score function S is standardized, i.e. $\mu = \sum_w S(w)p_v(w) = 0$ and $\sigma^2 = \sum_w S(w)^2 p_v(w) - \mu^2 = 1$, where $p_v(w) = 2^{-m}$ is the probability distribution of the inheritance vector under the *null hypothesis*, H_0: no linkage.

The NPL score for pedigree k can be written as

$$Z_k(x) = \sum_w P(v_k(x) = w)S_k(w), \tag{2.1}$$

where S_k is the score function and $v_k(x)$ is the inheritance vector at locus x for the kth pedigree, and $P(v_k(x) = w)$ the corresponding probability function for the inheritance vector given the marker data. With perfect marker information at locus x, we get $Z_k(x) = S_k(v(x))$.

For a collection of N pedigrees, the NPL score process (Kruglyak *et al.*, 1996; Kong and Cox, 1997; Teng and Siegmund, 1998) is defined as

$$Z(x) = \sum_{k=1}^{N} \gamma_k Z_k(x), \tag{2.2}$$

where $Z_k(x)$ is given in (2.1), γ_k is the weight assigned to the kth pedigree, and $\sum_{k=1}^N \gamma_k^2 = 1$. With full information from the markers, $E(Z(x)) = 0$ and $V(Z(x)) = 1$ under H_0. If the number of pedigrees is large, then $Z(x)$ will be marginally approximately $N(0, 1)$ distributed. This property is extensively used in the existing theory of linkage analysis. The weighting of the pedigree scores might be related to the pedigree size, information, structure, or the corresponding scores at different loci located on other

chromosomes. This last *conditional* multilocus approach has been used for the two-locus case by, for instance, Cox *et al.* (1999), Ängquist (2001), and Chiu and Liang (2004). The adequacy of the Normal approximation depends on the variability among the present pedigree structures and the set of weights.

Let Ω be the collection of markers at all chromosomes and define

$$Z_{\max} = \sup\{Z(x),\ x \in \Omega\}, \tag{2.3}$$

as the maximal NPL score among the markers.

2.3 *Significance of the results*

To interpret the significance of an observed maximal linkage score $z_{\max}$, we need to calculate the corresponding p-value

$$\alpha(z_{\max}) = P(Z_{\max} \geqslant z_{\max}|H_0). \tag{2.4}$$

If the number of pedigrees and markers and pedigree sizes are small enough, the genome-wide p-value in (2.4) may be computed exactly. Otherwise, one has to approximate the p-value using simulation techniques or theoretical approximations. Often, an assumption of Normally distributed linkage scores is made.

Note that $E(Z(x)|H_0) = 0$ but $V(Z(x)|H_0) < 1$, when the markers are not fully informative at x (in single-point or multipoint sense). This makes the p-value based on the *perfect data approximation* $V(Z(x)|H_0) = 1$ statistically conservative. We model imperfect marker data by assuming that the markers are equidistantly positioned (on all chromosomes) over Ω, with grid size Δ. However, all markers are assumed to be fully polymorphic, so that $V(Z(x)|H_0) = 1$ at all $x \in \Omega$. With this setup, perfect marker data corresponds to $\Delta = 0$.

An asymptotic approximation formula regarding the probability $\alpha(z_{\max})$ is given by Lander and Kruglyak (1995):

$$\alpha(z_{\max}) \approx 1 - \mathrm{e}^{-\mu(z_{\max})}, \tag{2.5}$$

where $\mu(z)$ is an approximation for the mean number of regions in which the linkage score process exceeds the threshold z. Lander and Kruglyak (1995) gave a formula for $\mu(z)$ when $\Delta = 0$ and $Z(x)$ is distributed as $N(0, 1)$ at all $x \in \Omega$. Here we will incorporate $\Delta > 0$ and possible skewness of Z, as suggested by Feingold *et al.* (1993) and Tang and Siegmund (2001). This gives the adjusted formula

$$\mu(z) = [C\kappa_1 + 2\rho\kappa_2\nu G z^2]\alpha_{\mathrm{pt}}(z), \tag{2.6}$$

where C is the number of chromosomes included in the scan and G the total genetic (in this case sex-averaged) length of these chromosomes. According to Collins *et al.* (1996), $G = 35.75$ Morgans if all the 22 autosomes are included in the scan. Moreover, the variable z refers to the score threshold and $\alpha_{\mathrm{pt}}(z) = 1 - \Phi(z)$ is the pointwise significance level of exceeding z when $Z(x)$ is standard Normal.

The *crossover rate* ρ is a function of the autocorrelation function $r_Z(\cdot)$ of the NPL process under H_0 and perfect marker data, and defined as

$$\rho = -\frac{r_Z'(0)}{2}, \tag{2.7}$$

where the derivative is taken from above. The crossover rate is a nonrandom quantity that measures the average amount of fluctuation of the NPL score statistic $Z(\cdot)$ and therefore depends both on the pedigree structures in the pedigree set as well as on the choice of score function. One example of a process satisfying (2.7) is the *Ornstein–Uhlenbeck* (*OU*)-process with mean 0 and covariance function $r_Z(\tau) = \mathrm{e}^{-2\rho|\tau|}$. However, the approximation formula (2.5) does not require $Z(\cdot)$ to be an OU-process.

The quantity $\nu = \nu(z(4\rho\Delta)^{1/2})$ adjusts for finite marker spacing. Here $\nu(\cdot)$ is the decreasing function defined in Siegmund (1985, p. 82). For perfect marker data, $\Delta = 0$, one has $\nu = 1$. Finally, the constants κ_1 and κ_2 adjust for possible skewness of $Z(x)$. When $E(Z(x)^3|H_0) = 0$, both κ_1 and κ_2 equal 1. See Dupuis and Siegmund (1999), Tu and Siegmund (1999), Tang and Siegmund (2001), and Section A of the Supplementary Material (henceforth denoted SM; see http://biostatistics.oupjournals.org/) for more details.

3. METHODS

3.1 *Calculating the crossover rate*

If using (2.7) as a starting point, it is possible to show that the following equality holds for a single pedigree of a general structure:

$$\rho = \frac{\lambda}{4} 2^{-m} \sum_{w} \sum_{j=1}^{m} (S(w) - S(w + e_j))^2, \tag{3.1}$$

where $\lambda = 1$ Morgan^{-1} or 0.01 centiMorgan^{-1} depending on which unit of genetic map distance is used. Moreover, w ranges over all 2^m binary vectors of length m and e_j is a binary vector with 1 in position j and zeros elsewhere. For more details, see Hössjer (2003a) and Section B of SM. This calculation is easily implemented in a computer program, for example, by using binary to decimal conversion for the different inheritance vectors and an index matrix reflecting all vector transformations $w \to w + e_j$.

Considering a pedigree set consisting of N pedigrees, it is possible to weight the different ρ-values into an overall value

$$\rho = \sum_{k=1}^{N} \gamma_k^2 \rho_k, \tag{3.2}$$

where ρ_k is the crossover rate for the kth pedigree.

3.2 *Calculating the crossover rate using Monte Carlo simulation*

The computational complexity $O(2^m)$ in (3.1) can be reduced to $O(2^{m-f})$ using founder phase symmetry (Kruglyak *et al.*, 1996; Gudbjartsson *et al.*, 2000) but, for large pedigrees, the computational complexity may still be burdensome. Under these circumstances, an option is to use a Monte Carlo approximation.

First notice that ρ may be reformulated as

$$\rho = \frac{\lambda}{4} E(f(v)), \tag{3.3}$$

where $f(w) = \sum_{j=1}^{m} (S(w + e_j) - S(w))^2$ and v is random with $P(v = w) = 2^{-m}$ for each w. A valid Monte Carlo approximation of ρ is therefore,

$$\hat{\rho} = \frac{\lambda}{4J} \sum_{i=1}^{J} f(v_i), \tag{3.4}$$

where $\{v_i\}_{i=1}^{J}$ are independent and identically distributed with the same distribution as v and J is the number of simulated inheritance vectors.

3.3 *Adjusting the approximation formula*

Formulas (2.5)–(2.6) to some extent correct for non-Normality by incorporating the skewness of Z via κ_1 and κ_2. In this section we describe how to fully adjust for non-Normality of $Z(x)$.

Let $F(z) = P(Z(x) \leqslant z|H_0)$ denote the marginal distribution function of $Z(\cdot)$ in (2.2). In our framework of fully polymorphic markers, $F(\cdot)$ is independent of $x \in \Omega$. In the work of Lander and Kruglyak (1995), the approximation $F = \Phi$ is assumed. A possible approximate correction of the distribution function, $F(\cdot)$, with respect to deviations from the standard Normal distribution, is to use Edgeworth expansions (see McCune and Gray, 1982) but in this case it is possible to calculate $F(\cdot)$ exactly with arbitrarily small error.

Define $Y(x) = g^{-1}(Z(x))$, where $g = (F^{-1} \circ \Phi)$. Since F is the distribution function of a discrete random variable, g does not have a unique inverse. We put $g^{-1} = (\Phi^{-1} \circ \widetilde{F})$, where $\widetilde{F}$ is the version of F which at points of discontinuity takes values halfway between the right-hand and left-hand limits, as shown in Section C of SM. Here Y is a stationary process, whose marginal distribution converges to Φ as F tends to a continuous function. It is possible to rewrite the unknown p-value as

$$\alpha(z) = P(Z_{\max} \geqslant z|H_0) = P(Y_{\max} \geqslant g^{-1}(z)|H_0). \tag{3.5}$$

In Section D of SM it is shown how $\rho_Y = -r_Y'(0)/2$ may be derived from $\rho = \rho_Z$. By combining (2.5) and (3.5) and replacing ρ with ρ_Y, we obtain

$$\alpha(z) \approx 1 - \mathrm{e}^{-\mu_{\mathrm{adj}}(z)}. \tag{3.6}$$

In (3.6)

$$\mu_{\mathrm{adj}}(z) = [C + 2\rho_Y \nu G g^{-1}(z)^2]\alpha_{\mathrm{pt}}(g^{-1}(z)), \tag{3.7}$$

where $\nu = \nu(z(4\rho_Y \Delta)^{1/2})$ adjusts for finite marker spacing. Compared to (2.6), notice that no skewness factors κ_1 and κ_2 are involved, since the transformation g removes skewness.

3.4 *Computing the p-value by Monte Carlo simulations*

An obvious way of approximating the p-value $\alpha(z)$ is to generate independent and identically distributed replicates $Z^1_{\max}, \ldots, Z^J_{\max}$ of $Z_{\max}$, under H_0, and then put

$$\alpha(z) \approx \frac{1}{J} \sum_{i=1}^{J} I(Z^i_{\max} \geqslant z), \tag{3.8}$$

where $I(A)$ is the indicator function for the event A. We will use the Monte Carlo approximation (3.8) in order to check the validity of the approximative p-value formulas (2.5) and (3.6).

We generate each $Z^i_{\max}$ by first simulating an NPL score process $Z^i(x)$ under H_0 and perfect marker information at *all* loci x, i.e. even between markers. The computational complexity is modest since no hidden Markov algorithm is needed for computing the appropriate pedigree scores. Assuming no chiasma interference (i.e. that the crossover points are independent of each other) and considering the inheritance on different chromosomes as independent, we simply progress the components of the inheritance vector as independent and stationary Markov processes with two states, 0 and 1, such that jumps (crossovers) between the states occur according to a Poisson process with intensity λ. Since each pedigree score is a deterministic function of the inheritance vector at all $x \in \Omega$, $Z^i_{\max}$ is easily computed from (2.2) to (2.3) once all inheritance vector processes have been simulated. Further results on simulation of linkage scores and p-values can be found in Boehnke (1986), Ploughman and Boehnke (1989), Ott (1989), Terwilliger *et al.* (1993), Malley *et al.* (2002), and Ängquist and Hössjer (2004).

4. RESULTS

We have applied the theory from the preceding section to different kinds of pedigree sets. All computer calculations have been performed using MATLAB. Throughout, we let Ω represent all the 22 human autosomes with a total sex-averaged autosomal genome length G of 35.75 Morgans (Collins *et al.*, 1996), and use equal weighting for all the N pedigrees included in the pedigree set (i.e. $\gamma_k = 1/\sqrt{N}$).

4.1 *Calculating the significance using the crossover rate*

We calculated the crossover rate using (3.1) for pedigrees of the seven different structures shown in Figure 1. The results are shown in Table 1. Pedigrees 1–4 replicate the results given by Lander and Kruglyak (1995). Pedigrees 6–7 are both present in the second BOTNIA study, which is further described in Section 4.2.

As pointed out above, the ρ-value measures the fluctuation rate of the NPL process $Z(x)$. This means, loosely speaking, that a high value of ρ will make it easier for the process to attain extreme values and therefore, with a higher probability, somewhere on the corresponding genome region exceed a given threshold z. The p-value is therefore an increasing function of ρ. In addition, note that the crossover rate depends on the choice of score function.

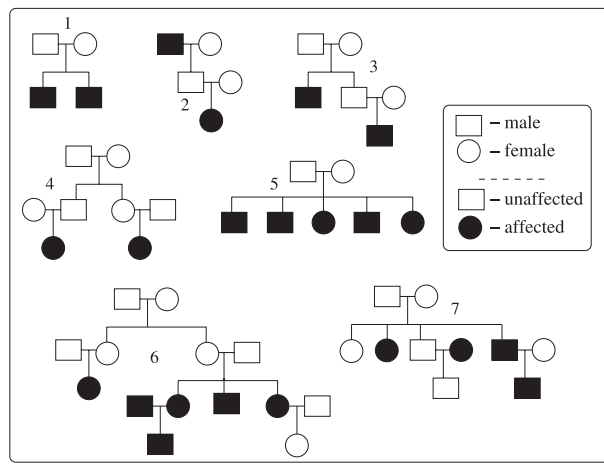

Fig. 1. Pedigrees corresponding to the ρ-values given in Table 1.

Table 1. *Crossover rate (in Morgans^{-1}) for pedigrees of different structures (cf. Figure 1)*

Example	Pedigree structure	ρ-value (S_{all})
1	Sib-pairs	2.0000
2	Grandparent/grandchild	1.0000
3	Uncle/nephew	2.5000
4	First cousins	2.6667
5	Five affected siblings	2.0847
6	BOTNIA pedigree	2.5053
7	BOTNIA pedigree	2.1880

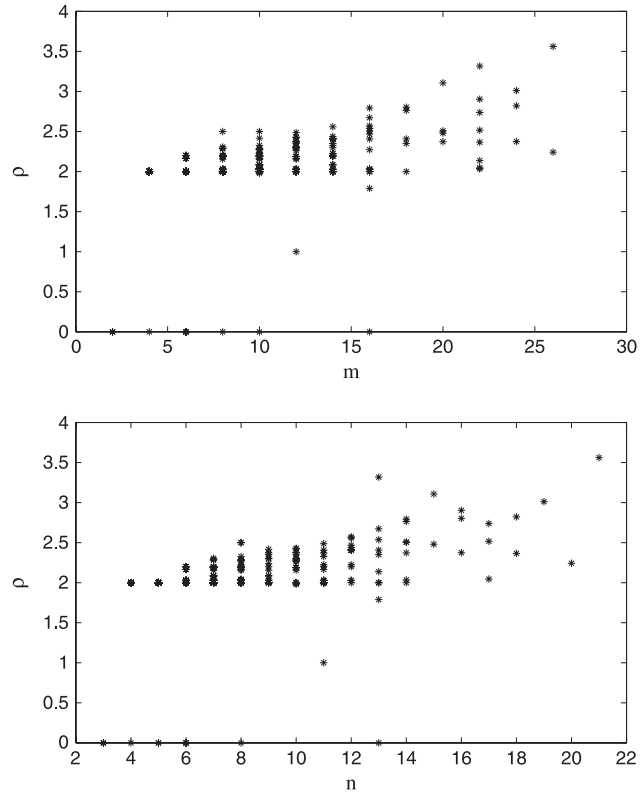

Fig. 2. Plotting the crossover rate (ρ) against the number of meioses (m) and individuals (n).

Table 2. *Crossover rate for the BOTNIA pedigree set—exact or using Monte Carlo approximation (cf. Figure 2)*

Type of calculation	Number of pedigrees	$\rho_{\max}$	$\rho_{\min,\ \rho>0}$	$\rho_{\min}$	Mean (ρ)	Std (ρ)
Exact	324	3.1069	1.0000	0.0000	2.0246	0.4031
Monte Carlo	13	3.5620	2.0348	2.0348	2.6208	0.4868
Total	337	3.5620	1.0000	0.0000	2.0476	0.4218

4.2 *Further properties of the crossover rate*

To get a clearer view of the distribution of the crossover rate, we calculated the parameter ρ for the pedigree set from the second BOTNIA type-2-diabetes study (Parker *et al.*, 2001; Lindgren *et al.*, 2002). This set consisted of 337 different pedigrees, originating from Finland and Sweden. Exact calculations were performed for the 324 pedigrees with $m \leqslant 20$ meioses. The remaining crossover rates (13 pedigrees) were estimated using the Monte Carlo simulation technique (3.4) with $J = 1000$. A summary of these outcomes is given in Table 2 and Figure 2, where plots of ρ against both m (number of meioses) and n (number of individuals) are shown. There is some positive correlation between ρ and m (0.4115) and ρ and n (0.4140), whereas m and n are highly correlated (0.9738).

Moreover, when the value of the score function is independent of the inheritance vector, $\rho = 0$. This follows from using (3.1) or the formulas in Section B of SM. An example is a pedigree consisting only of two unaffected parents and one affected child.

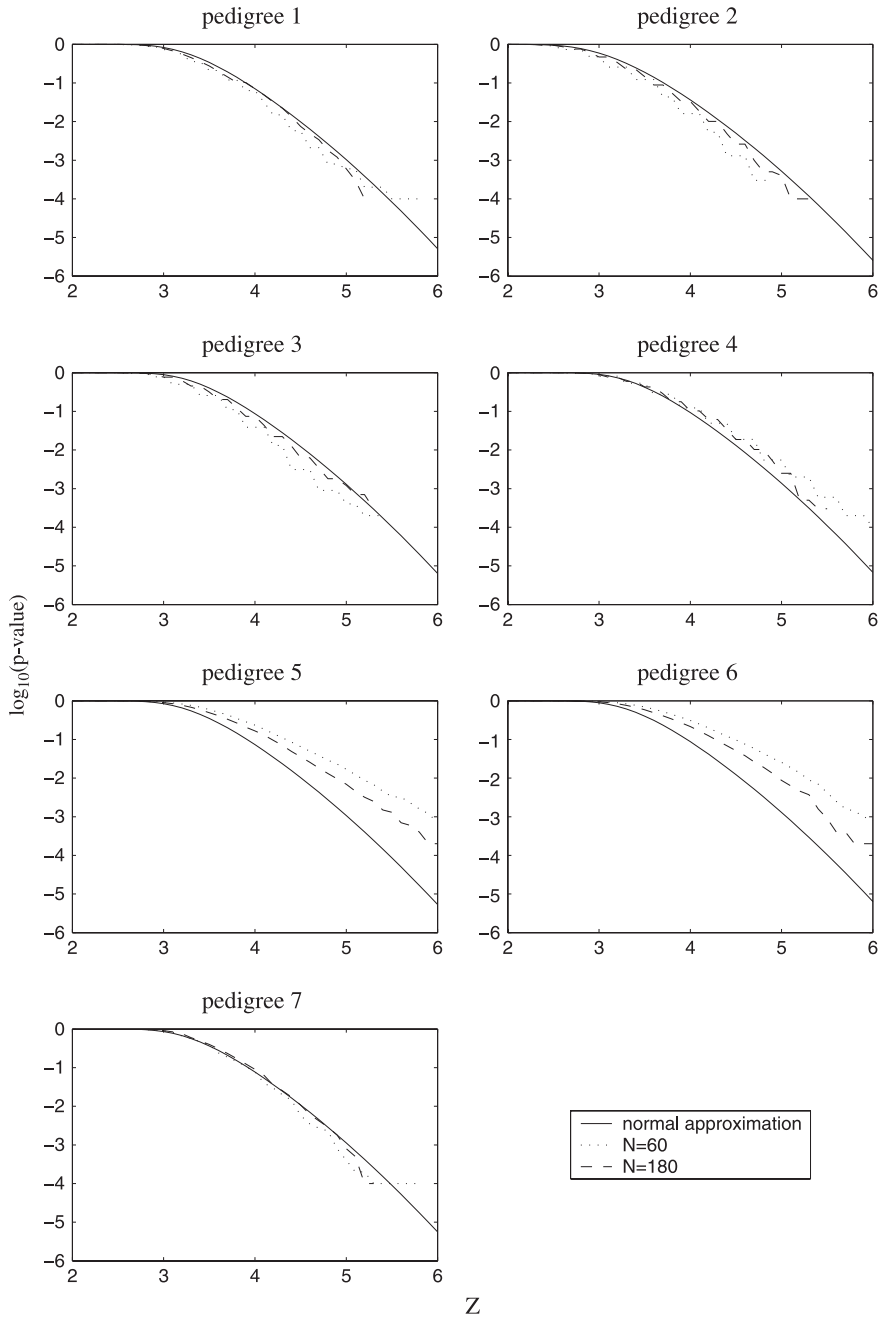

Fig. 3. Comparisons between the p-values for the Normal (unadjusted) approximation (2.5) and the simulation procedure (3.8) with $\Delta = 0$, $J = 10\,000$, and $N = 60$, 180.

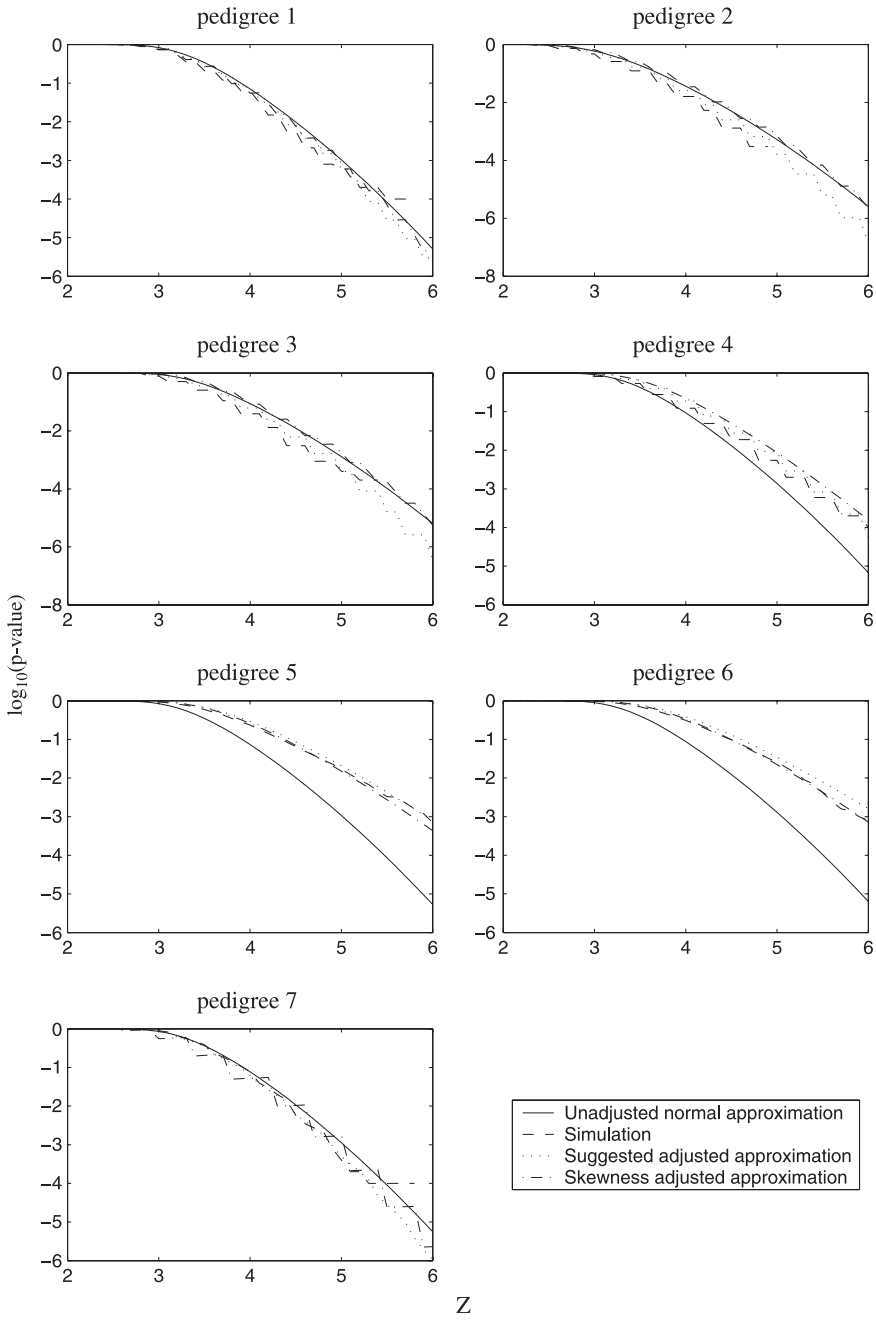

Fig. 4. Comparisons between the p-values for the unadjusted and skewness adjusted approximations in (2.5), the simulation procedure (3.8), and our non-Normality adjusted approximation (3.6). Further, $\Delta = 0$, $J = 10\,000$, $N = 60$, and the accuracy parameter for the probability distribution approximation is set to $\epsilon = 0.001$ and $l = 5$ (cf. Sections C–D in SM).

4.3 *Calculating the p-value using simulations*

We applied the simulation to the seven pedigrees displayed in Figure 1. The marker map was infinitely dense ($\Delta = 0$) and the number of simulations $J = 10\,000$. We compared the performance of the asymptotic approximation formulas (2.5)–(2.6), with no skewness correction ($\kappa_1 = \kappa_2 = 1$) and ρ as in (3.1)–(3.2), to the simulation results. To see more clearly the effect of the pedigree structure on the p-value, we considered homogeneous pedigree sets (i.e. all pedigrees identical) of sizes $N = 60$ and $N = 180$ for all seven examples. The results are shown in Figure 3 as p-values for thresholds 2.0, 2.1, …, 6.0.

The distribution of the simulated process depends on the number of pedigrees, their pedigree structure, and the choice of scoring function. The performance of the unadjusted approximation formula depends on how well the discrete distribution of Z is approximated by the standard Normal distribution (cf. Section 4.4) and, by the *central limit theorem*, on the number of pedigrees.

4.4 *Calculating the p-value using the adjusted approximation formula*

We tested the accuracy of the adjusted p-value approximation (3.6), again using the seven pedigree structures from Figure 1 to construct homogenous pedigree sets with $N = 60$. Taking $\Delta = 0$, we compared the original asymptotic approximation formula (with and without skewness correction), the Monte Carlo simulation technique, and our suggested approximation formula. The results are given in Figure 4. They support the view that it is important to adjust for non-Normality (especially skewness), since (3.6) and the skewness adjusted approximation formula clearly outperform the nonadjusted version of (2.5). The best choice of approximation seems to depend on the pedigree set, as will be discussed in Section 4.6.

In the unadjusted formula, the distribution of Z is approximated by a standard Normal distribution. Loosely speaking, skewed distributions with a thick or narrow right-hand tail will give anticonservative or conservative results, respectively, for p-values corresponding to the upper tail of F. Moreover, symmetric distributions with truncated tails, for instance because of very few meioses m, will give conservative p-values for large thresholds z. This is formalized by using the theory of Edgeworth expansions. In our case we calculated the first four *cumulants* (skewness and kurtosis) of the distributions (cf. Table 3 and Figure 5)

$$\begin{aligned} F_{\text{diff}}(z) &= F(z) - \Phi(z) \\ &\approx -\frac{k_3}{6}(z^2-1)\Phi'(z) - \frac{k_4}{4}(z^3-3z)\Phi'(z) - \frac{k_3^2}{72}(z^5-10z^3+15z)\Phi'(z), \end{aligned} \tag{4.9}$$

where k_3 and k_4 are the third and fourth cumulants, respectively, and F_{diff} measures the numerical difference, at value z, between the distribution of Z and a standard Normal variable.

Table 3. *The first four cumulants for pedigrees* 1–7

Pedigrees	k_1	k_2	k_3	k_4
1	0	1	0	−1
2–3	0	1	0	−2
4	0	1	1.1547	−0.6667
5	0	1	1.9972	4.1044
6	0	1	2.2722	6.6118
7	0	1	0	−1.64

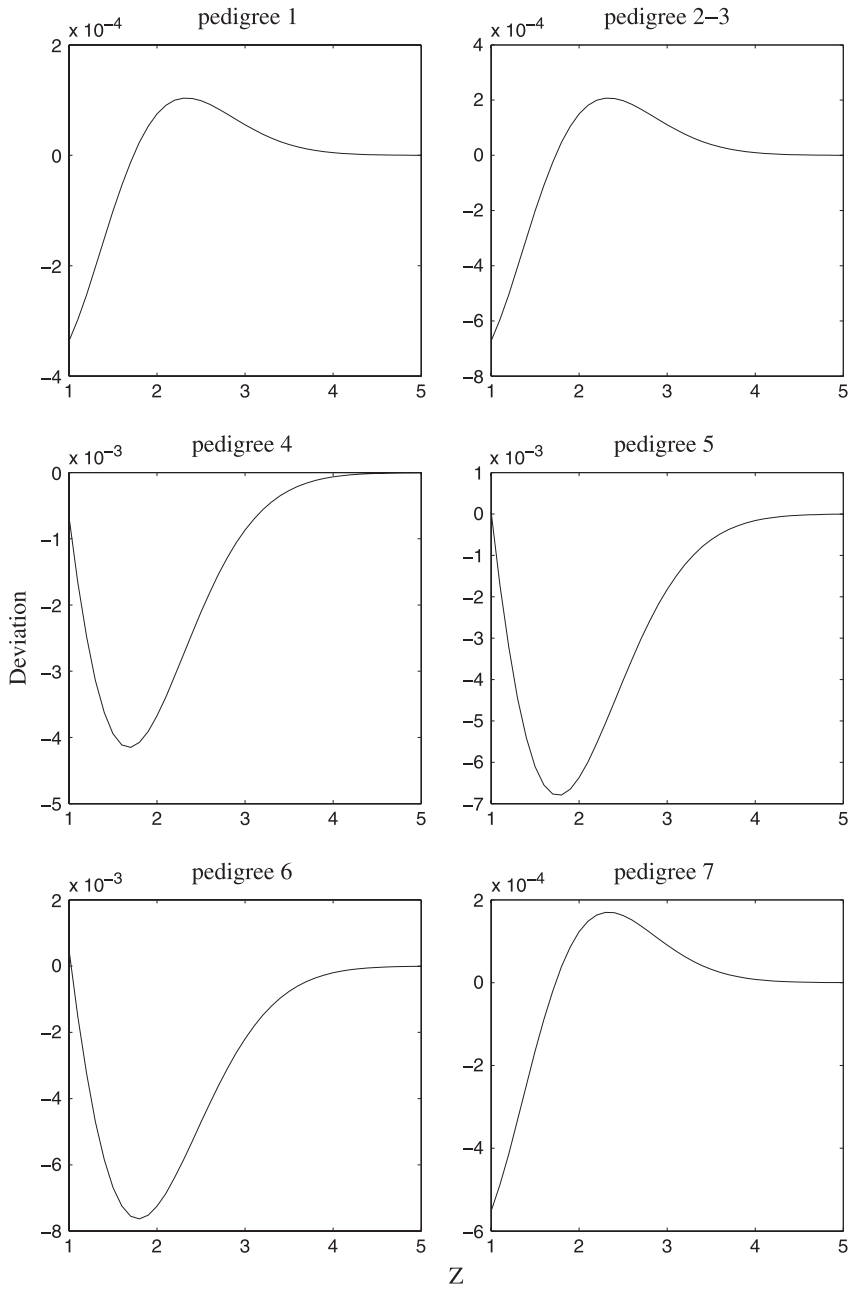

Fig. 5. The difference $F - \Phi$ between F and the cumulative distribution function Φ for the standard Normal distribution using Edgeworth expansions and $N = 60$.

The Edgeworth expansions explain the bias of the p-value approximations based on the assumption of Normal marginal distributions. Pedigrees 1–3 and 7 have positive deviations $F_{\text{diff}}(z)$ from the Normal distribution, for large z, and the original asymptotic approximation formula is therefore conservative. On the contrary, $F_{\text{diff}}(z)$ is negative for large z for pedigrees 4–6, implying anticonservativeness. Further,

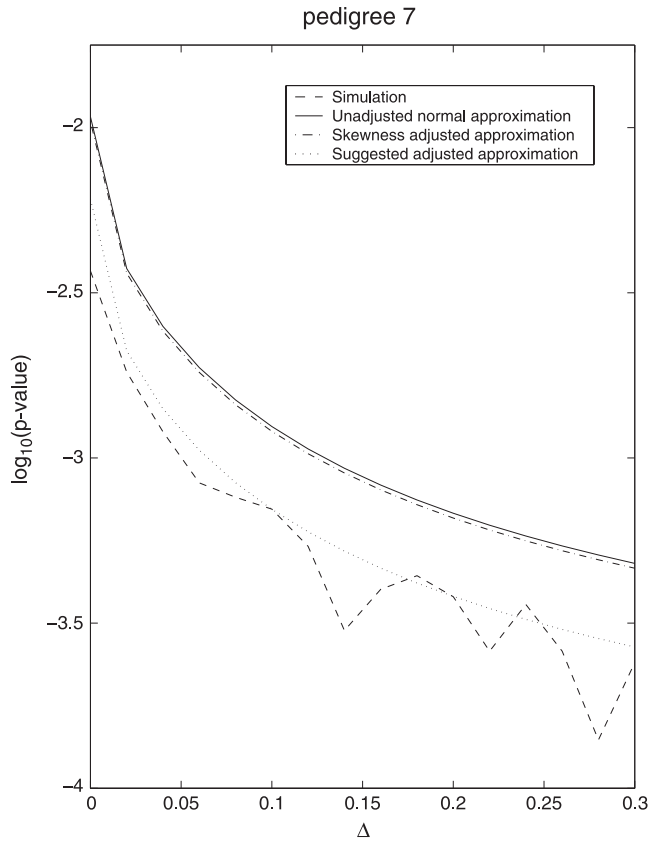

Fig. 6. Estimated p-values with respect to marker distance $\Delta = 0 : 0.02 : 0.30$ Morgans for threshold $T = 4.5$, using $J = 50\,000$ simulations. For further assumptions cf. Figure 4.

the adjusted approximation formulas seem to perform better than the unadjusted formula in all of the seven cases.

In Figure 6 we illustrate, using $N = 60$ and $T = 4.5$, the amount of decrease of the p-value approximations with respect to increased equidistant marker spacing Δ Morgans. One may note that generally the variance of the simulation estimator increases with larger Δ since more simulations are needed to reach a reasonable number of exceedances of T.

4.5 *Pedigree structures from two real examples*

We applied the new method to two pedigree sets from distinct studies. Firstly, we used the pedigree set from the BOTNIA study (cf. Section 4.2). In this case we selected only the 266 pedigrees with both the number of meioses $m \leqslant 12$, to reduce the computational complexity, and a nonconstant score function, hence $\rho > 0$. These pedigrees had crossover rates in the interval [1.0; 2.4309] and the corresponding mean value was 2.0390. The p-value estimates are displayed in Figure 7. The previous interpretations (cf. Section 4.4) seem still to hold as the results confirm, with satisfactory accuracy, the corresponding theory. An Edgeworth expansion, not shown, gives negative deviation F_{diff}, implying anticonservative tests.

Secondly, we used the pedigree set shown in Figure 8, consisting of only one single large pedigree with a very skewed distribution F. Using the notation and theory from Section D of SM, we found that

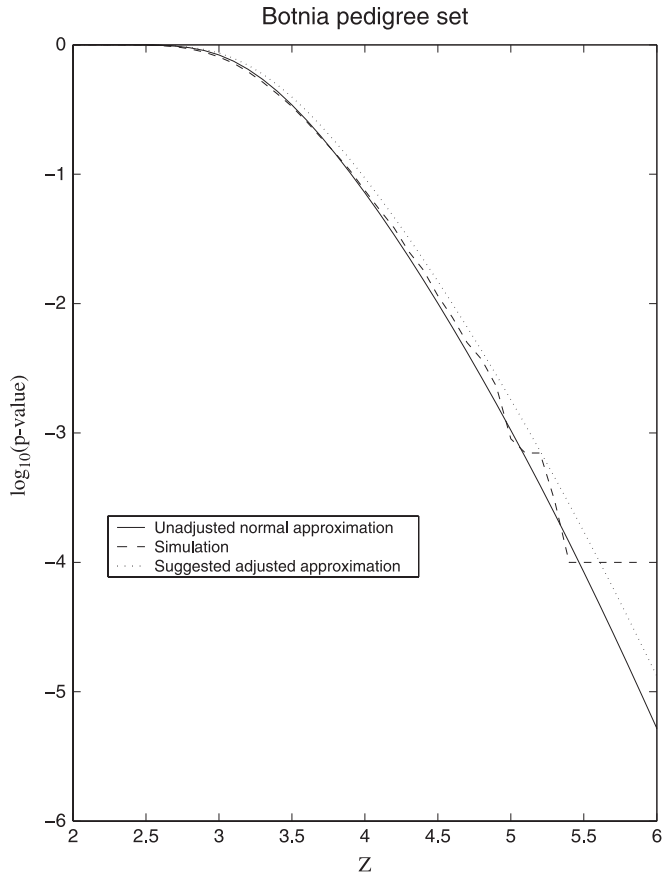

Fig. 7. Comparisons, for the BOTNIA pedigree set, between the p-values for the unadjusted approximation in (2.5), the simulation procedure (3.8), and our non-Normality adjusted approximation (3.6). Further, $\epsilon = 0.0005$ and $J = 10\,000$.

$\sum_{k=1}^{5} \alpha_k^2 = 0.8795$ which is far from 1 and therefore contradicts (D.2). In order to not obtain very conservative significance results using (D.3), we increased l, the number of influential Hermite polynomials, to 100 with $\sum_{k=1}^{100} \alpha_k^2 = 0.9963$. Moreover, we calculated $\rho = 4.1288$ and $\sum_{k=1}^{100} k\alpha_k^2 = 8.9602$ which gave us a final crossover rate of $\rho_Y = 0.4608$. This pedigree illustrates the importance of adjusting for non-Normality. Figure 9 shows the results. The very good performance of our adjusted approximation formula compared to (2.5) in this case probably depends on the large difference between F and Φ.

4.6 *Properties and performance*

It is clear that the adjustment for non-Normality (3.6)–(3.7) will be of increasing importance the more F departs from Φ whereas, as appears to be the case for the BOTNIA pedigree set, when F is close to Normal, the two methods are virtually equal.

The results in Figures 4, 6, 7, and 9 indicate that formula (3.6) is slightly conservative, at least when $\Delta = 0$. There may be several reasons for this. First, one may choose the parameter l too small leading to a too large ρ_Y in (D.4) of SM, which in turn implies that the p-values are overestimated. Secondly, the original formulas (2.5)–(2.6) are most accurate for OU-processes. However, the true covariance function

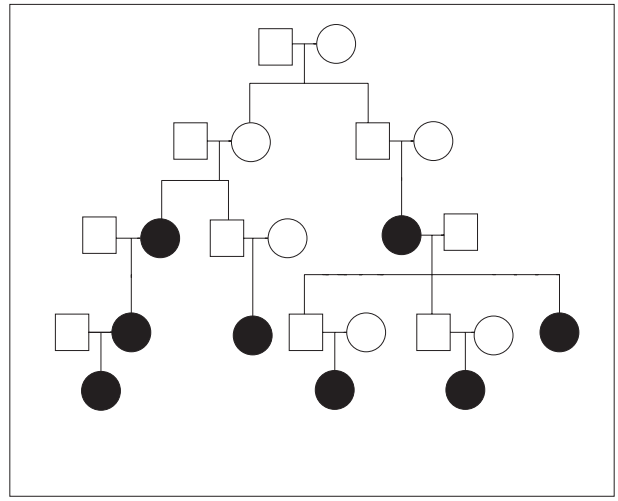

Fig. 8. The pedigree structure corresponding to the second real pedigree set example, cf. Figure 7.

of Z in general differs from $e^{-2\rho|t|}$ (Hössjer, 2003c), and this may cause some conservativeness of (2.5)–(2.6) and (3.6)–(3.7). A third possible source of conservative p-values may be the choice of clumping rate and the $L = 0$ correction in (A.1) of SM.

Except for this small conservative bias, the suggested approximation formula often seems to be more accurate than the skewness adjusted version of (2.5), probably because we not only correct for skewness but also transform F to a standard Normal distribution. Both methods seem to have good performance in general and the best choice depends on the pedigree set.

When $N = 1$, it is possible to use another method when calculating the adjusted crossover rate ρ_Y. This approach is based on the substitution of Z with $g^{-1}(Z)$ in the original crossover formula (3.1). In the second real data example, this method suggests a substantially larger value of ρ_Y, but the performance of the p-value estimation is in this case not as accurate as when using the first method. In other cases this method may be preferable.

Further discussion of related issues, e.g. concerning distributions and their relation to skewness, conservativeness, and deviations from the standard Normal distribution, may be found in Kong and Cox (1997), Teng and Siegmund (1998), Nicolae *et al.* (1998), Sengul *et al.* (2001), and Tang and Siegmund (2001). In these articles some of the problems mentioned above are, for instance, discussed in the context of different scoring functions and pedigree structures (mainly nuclear families).

5. DISCUSSION

5.1 *General comments*

In this article we have described two improvements of significance calculation for genome-wide scans through the approximation formula given in Lander and Kruglyak (1995) and extended by Feingold *et al.* (1993) and Tang and Siegmund (2001).

Firstly, a method of calculating ρ for an arbitrary pedigree set is presented. Secondly, in the unadjusted formula an assumption of Normally distributed NPL scores is made. If this assumption fails, the estimated p-values might be either conservative or anticonservative, to a large extent depending on the cumulants of order 3 and 4 of the marginal distribution function F of the NPL score. In this work we present

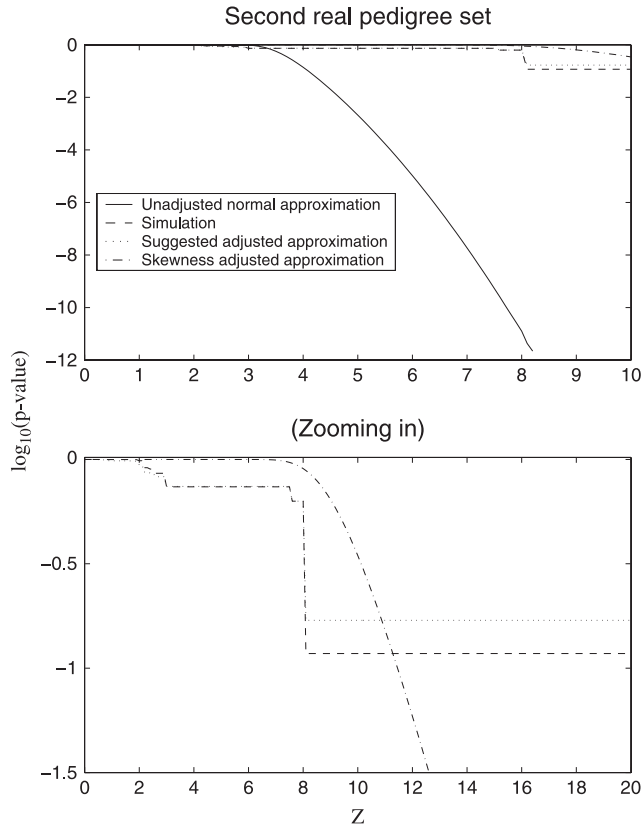

Fig. 9. Comparisons, for the second real pedigree set example, between the p-values for the unadjusted and skewness adjusted approximations in (2.5), the simulation procedure (3.8), and our non-Normality adjusted approximation (3.6). Moreover, $\epsilon = 0.001$ and $J = 100\,000$. The two largest values, the standardized score function S_{all} can attain are 31.6115 and 8.0830. The largest value corresponds to the case where all the affected individuals share an allele *IBD*, and this possibility is the reason behind the extreme skewness of F in this case.

an approach which takes the pedigree set structure into account in such a way that the p-values will be corrected for deviations from Normality. We compared the performance of the original, both the unadjusted and the skewness adjusted formulas, with our non-Normality adjusted formula and found that the latter seems to be successful, not only because of the robust performance but also since the p-value curve, as a function of thresholds, more closely follows the discrete behavior of the NPL score. The use of the non-Normality adjusted formula (3.6) also forces us to update the crossover rate (3.7). This correction may be described using a Hermite polynomial expansion of the transformation g between the standard Normal distribution Φ and the marginal distribution F of the NPL score, see Section D of SM.

5.2 *Assumptions and further work*

Throughout this article we have made some assumptions to gain simplicity as follows. Firstly, we do not discriminate the genetic genome lengths with respect to genders and therefore use sex-averaged values. One may note that there is no theoretical problem in generalizing (3.1) to sex-specific genetic lengths. If we assume λ_f and λ_m to equal the genetic map length in Morgans per unit of measurement (x) for females

and males, respectively, we get

$$\rho = \frac{1}{4} 2^{-m} \sum_{w} \sum_{j=1}^{m} \lambda_j (S(w) - S(w + e_j))^2, \tag{5.1}$$

where λ_j equals λ_f or λ_m depending on whether the jth meioses corresponds to the formation of an egg or sperm cell. Secondly, the crossover rate formula (3.1) is valid for several types of map functions (Hössjer, 2003a), although we used Haldane's map function in the simulations. This implies, for instance, that the absence of chiasma interference is assumed. Thirdly, the markers are assumed equidistant and fully polymorphic. The effects of all these assumptions deserve further study.

A partial solution in the case of incomplete marker information may be to use a transformation (as in Section 3.3) to the $1 - \mathrm{df}$ likelihood ratio statistic Z_{lr} of Kong and Cox (1997) and Nicolae *et al.* (1998). This statistic is often less skewed than Z, in which case the transformation g might be closer to the identity function, making the crossover adjustment of Section D in SM less crucial.

ACKNOWLEDGMENTS

This research is sponsored by the Swedish Research Council, under contracts 6152-8013 and 621-2001-3288. Financial support is also given by the Wallenberg Laboratory, Department of Endocrinology, Malmö University Hospital, Lund University, Malmö. We wish to thank professor Leif C. Groop, Wallenberg Laboratory, for the permission to use the BOTNIA data set in the analyses. The authors would also like to thank Professor Peter J. Diggle and anonymous referees for valuable suggestions of how to improve the article.

REFERENCES

ABECASIS, G. R., CHERNY, S. S., COOKSON, W. O. AND CARDON, L. R. (2002). MERLIN—rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* **30**, 97–101.

ABNEY, M., OBER, C. AND MCPEEK, M. S. (2002). Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the Hutterites. *American Journal of Human Genetics* **70**, 920–934.

ÄNGQUIST, L. (2001). Conditional two-locus NPL-analyses: theory and applications. *Master's thesis No. 2001:E22*, Lund University, Department of Mathematical Statistics, Lund.

ÄNGQUIST, L. AND HÖSSJER, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications in Genetics and Molecular Biology* **3**, 24 pp. (electronic journal).

BOEHNKE, M. (1986). Estimating the power of a proposed linkage study: a practical computer simulation approach. *American Journal of Human Genetics* **39**, 513–527.

CHIU, Y. F. AND LIANG, K. Y. (2004). Conditional multipoint linkage analysis using affected sib pairs: an alternative approach. *Genetic Epidemiology* **26**, 108–115.

COLLINS, A., FREZAL, J., TEAGUE, J. AND MORTON, N. E. (1996). A metric map of humans: 23.500 loci in 850 bands. *Proceedings of the National Academy of Sciences of the United States of America* **93**, 14771–14775.

COX, N. J., FRIGGE, M., NICOLAE, D. L., CONCANNON, P., HANIS, C. L., BELL, G. I. AND KONG, A. (1999). Loci on chromosomes 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. *Nature Genetics* **21**, 213–215.

DAVIS, S. AND WEEKS, D. E. (1997). Comparisons of nonparametric statistics for detection of linkage in nuclear families: single-marker evaluations. *American Journal of Human Genetics* **61**, 1431–1444.

DOERGE, R. W. AND CHURCHILL, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

DUPUIS, J. AND SIEGMUND, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. *Genetics* **151**, 373–386.

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

FEINGOLD, E., SONG, K. K. AND WEEKS, D. E. (2000). Comparisons of allele-sharing statistics for general pedigrees. *Genetic Epidemiology* **19** (Suppl. 1), 92–98.

GUDBJARTSSON, D. F., JONASSON, K., FRIGGE, M. AND KONG, A. (2000). ALLEGRO, a new computer program for multipoint linkage analysis. *Nature Genetics* **25**, 12–13.

HÖSSJER, O. (2003a). Asymptotic estimation theory of multipoint linkage analysis under perfect marker information. *Annals of Statistics* **31**, 1075–1109.

HÖSSJER, O. (2003b). Determining inheritance distributions via stochastic penetrances. *Journal of the American Statistical Association* **98**, 1035–1051.

HÖSSJER, O. (2003c). Spectral decomposition of score functions in linkage analysis. *Technical Report No. 2003:21*. Stockholm: Stockholm University, Department of Mathematical Statistics.

HÖSSJER, O. (2005). Conditional likelihood score functions for mixed models in linkage analysis. *Biostatistics* **6**, 313–332.

KONG, A. AND COX, N. (1997). Allele-sharing models: Lod scores and accurate linkage tests. *American Journal of Human Genetics* **61**, 1179–1188.

KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. AND LANDER, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American Journal of Human Genetics* **55**, 1347–1363.

KRUGLYAK, L. AND LANDER, E. S. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

LANDER, E. S. AND BOTSTEIN, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

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

LINDGREN, C. M., MAHTANI, M. M., WIDÉN, E., MCCARTHY, M. I., DALY, M. J., KIRBY, A., REEVE, M. P., KRUGLYAK, L., PARKER, A., MEYER, J., ALMGREN, P., LEHTO, M., KANNINEN, T., TUOMI, T., GROOP, L. C. AND LANDER, E. S. (2002). Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the BOTNIA study. *American Journal of Human Genetics* **70**, 509–516.

MALLEY, J. D., NAIMAN, D. AND BAILEY-WILSON, J. (2002). A comprehensive method for genome scans. *Human Heredity* **54**, 174–185.

MCCUNE, E. D. AND GRAY, H. L. (1982). Cornish–Fisher and Edgeworth expansions. In Kotz, S. and Johnson, N. L. (eds), *Encyclopedia in Statistical Sciences*, Volume 2. New York: John Wiley & Sons, pp. 188–193.

MCPEEK, M. S. (1999). Optimal allele-sharing statistics for genetic mapping using affected relatives. *Genetic Epidemiology* **16**, 225–249.

NICOLAE, D. L. (1999). Allele sharing models in gene mapping: a likelihood approach, *Doctoral thesis*, University of Chicago, Department of Statistics, Chicago, IL.

NICOLAE, D. L., FRIGGE, M. L., COX, N. J. AND KONG, A. (1998). Discussion. *Biometrics* **54**, 1271–1274. (Discussion of article by Teng and Siegmund, 1998.)

OTT, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America* **86**, 4175–4178.

OTT, J. (1999). *Analysis of Human Genetic Linkage*, 3rd edition. New York: The John Hopkins University Press.

PARKER, A., MEYER, J., LEWITZKY, S., RENNICH, J. S., CHAN, G., THOMAS, J. D., ORHO-MELANDER, M., LEHTOVIRTA, M., FORSBLOM, C., HYRKKÖ, A., CARLSSON, M., LINDGREN, C. AND GROOP, L. C. (2001). A gene conferring susceptibility to type 2 diabetes in conjunction with obesity is located on chromosome 18p11. *Diabetes* **50**, 675–680.

PLOUGHMAN, L. M. AND BOEHNKE, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44**, 543–551.

SENGUL, H., WEEKS, D. E. AND FEINGOLD, E. (2001). A survey of affected-sibship statistics for nonparametric linkage analysis. *American Journal of Human Genetics* **69**, 179–190.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals* [Springer Series in Statistics]. Berlin: Springer.

TANG, H. K. AND SIEGMUND, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics* **2**, 147–162.

TENG, J. AND SIEGMUND, D. (1998). Multipoint linkage analysis using affected relative pairs and partially informative markers. *Biometrics* **54**, 1247–1265.

TERWILLIGER, J. D., SPEER, M. AND OTT, J. (1993). Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genetic Epidemiology* **10**, 217–224.

TU, I. P. AND SIEGMUND, D. (1999). The maximum of a function of a Markov chain and applications to linkage analysis. *Advances in Applied Probability* **31**, 510–531.

WHITTEMORE, A. S. (1996). Genome scanning for linkage: an overview. *American Journal of Human Genetics* **59**, 704–716.

WHITTEMORE, A. S. AND HALPERN, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics* **50**, 118–127.

[*Received April 20, 2004; first revision January 15, 2005; second revision February 11, 2005; accepted for publication March 7, 2005*]