Assessing Genomewide Statistical Significance in Linkage Studies

D.Y. Lin* **and Fei Zou**

Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina

Assessment of genomewide statistical significance in multipoint linkage analysis is a thorny problem. The existing analytical solutions rely on strong assumptions (i.e., infinitely dense or equally spaced genetic markers that are fully informative and completely observed, and a single type of relative pair) which are rarely satisfied in real human studies, while simulation-based methods are computationally intensive and may not be applicable to complex data structures and sophisticated genetic models. Here, we propose a conceptually simple and numerically efficient Monte Carlo procedure for determining genomewide significance levels that is applicable to all linkage studies. The pedigree structure is completely general; the marker data are totally arbitrary in respect to number, spacing, informativeness, and missingness; the trait can be qualitative, quantitative, or multivariate; the alternative hypothesis can be two-sided or one-sided; and the statistic can be parametric or nonparametric. The usefulness of the proposed approach is demonstrated through extensive simulation studies and an application to the nuclear family data from the Tenth Genetic Analysis Workshop. © 2004 Wiley-Liss, Inc.

Key words: dense map; Gaussian process; genome scan; Haseman-Elston regression; IBD; LOD score; Monte Carlo; multipoint linkage analysis; sparse map; thresholds

Grant sponsor: National Cancer Institute

*Correspondence to: Danyu Lin, Ph.D., Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420.
E-mail: lin@bios.unc.edu

Received 23 March 2004; Accepted 5 May 2004
Published online 13 July 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI: 10.1002/gepi.20017

INTRODUCTION

Linkage analysis aims to extract available inheritance information from pedigree data and to test for coinheritance of chromosomal regions with a trait. Traditional linkage analysis is based on studying individual genetic markers one at a time, and thus does not make full use of inheritance information. Spurred by high-resolution genetic maps and automated genotyping technology, together with the ground-breaking work of Kruglyak and Lander [1995a] and Kruglyak et al. [1996] among others, it has become a common practice to perform multipoint linkage analysis, which examines a large number of genetic markers simultaneously so as to assess the presence of trait-influencing genes in a comprehensive and efficient manner.

Multipoint or genomewide linkage studies consist of three steps: scan the entire genome with a collection of genetic markers; calculate an appropriate linkage statistic at each position along the genome; and identify the regions in which the statistic shows a significant deviation from what would be expected under independent assortment [Lander and Kruglyak, 1995]. Since the study aims to detect a deviation somewhere in the whole genome, the significance level pertains to the probability that the maximum of the test statistic over the whole genome exceeds a certain threshold value. The determination of threshold values for such maximum deviation tests turns out to be a highly challenging statistical problem.

Elegant analytical solutions under idealized conditions have been derived for quantitative trait loci (QTL) mapping in experimental crosses [Lander and Botstein, 1989; Kruglyak and Lander, 1995b; Dupuis and Siegmund, 1999; Zou et al., 2001] and for affected-relative-pair analysis in humans [Feingold et al., 1993]. These analytical results require that the markers are infinitely dense and fully informative, and that recombinations occur without interference. Approximations have been developed for discrete sets of fully informative markers under the assumptions that the markers are equally spaced with no missing values, and that the linkage statistic is evaluated at the markers only [Feingold et al., 1993; Dupuis and Siegmund, 1999; Zou et al., 2001]. The analytical results for the affected-relative-pair analysis further require that the study involves only one class of relative pair, although approximations have been proposed for a mixture of relative pair [Feingold et al., 1993; Lander and Kruglyak, 1995]. It was suggested that the threshold for the affected grandparent-grandchild analysis be applied to the LOD score analysis of individual informative meioses in nuclear families [Lander and Schork, 1994; Lander and Kruglyak, 1995], although no formal derivations have been given. As pointed out by Lander and Schork [1994], the assumptions required by the analytical approach "will often be false in important situations." No analytical results are available for general human pedigrees, especially with quantitative or multivariate traits.

Permutation tests have been proposed to determine the thresholds for QTL mapping [Churchill and Doerge, 1994; Doerge and Churchill, 1996] and for human sib pairs [Wan et al., 1997]. This approach has the advantage of making no assumptions on the distribution of the phenotype. However, it is computationally intensive, since the analysis needs to be repeated for each permuted data set, and it is not applicable to certain study designs/pedigree structures (e.g., pedigrees consisting of affected sib pairs and their parents). A number of authors [e.g., Ott, 1989; Weeks et al., 1990; Davis et al., 1996; Sawcer et al., 1997; Guerra et al., 1999] proposed approximating the significance levels by simulating the genotypes of each individual in the pedigrees. This approach is very time-consuming, especially for complex pedigrees with missing data. Zhao et al. [1999] proposed a more efficient simulation procedure, but the method may be biased when there are missing genotypes or when the pedigrees involve three or more generations. No matter how the genotypes are simulated, the analysis needs to be repeated for each simulated data set, which is time-consuming. Recently, Song et al. [2004] proposed an algorithm to reduce the number of simulated data sets. These simulation methods cannot be used if one is interested in locating a gene given other known genes on the same chromosome.

In this article, we develop a simple and efficient Monte Carlo approach to determining genomewide significance levels that is applicable to all linkage statistics. We allow completely general pedigree structures, and accommodate qualitative, quantitative, and even multivariate traits. The markers are arbitrarily spaced and are allowed to have arbitrary degrees of informativeness and missingness. We show that, under the null hypothesis of no linkage, the distributions of the linkage statistics are functions of certain zero-mean Gaussian processes indexed by the genome location, whose realizations can be approximated by Monte Carlo simulation. We can then use the Monte Carlo distributions to approximate the thresholds or significance levels of the desired maximum deviation statistics. This approach has much broader applications than the currently available approaches, and is computationally much less demanding than the use of permutation and other simulation methods. Extensive simulation studies and an application to the nuclear family data from the Tenth Genetic Analysis Workshop reveal that: 1) the proposed approach yields thresholds that are close to the true values and thus maintain proper type I error and power; 2) the analytical thresholds based on the dense-marker theory tend to be overly conservative and thus incur loss of power, even when the markers are as dense as 1 cM apart; and 3) the proposed approach enables one to make more efficient use of the data than the analytical approach in important situations.

METHODS

DEFINITIONS

Statistical tests in genetic linkage studies can be formulated in terms of testing the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta \in C$ in the presence of a nuisance parameter γ, where the alternative parameter space C depends on the specific problem; β or γ or both may consist of more than one component. For QTL mapping in experimental crosses [Lander and Botstein, 1989; Zeng, 1994; Jansen and Stam, 1994; Kao and Zeng, 1997], β pertains to the genetic effects of the putative QTL, while γ corresponds to the grand mean, the error variance, and possibly the effects of covariates and other QTLs. In this case, we have the familiar two-sided alternatives $H_1 : \beta \neq 0$. For the classical linkage analysis, we set $\beta = f - 1/2$, where f is the recombination fraction, and test $H_0 : \beta = 0$ against $H_1 : \beta < 0$. For the variance-components analysis of quantitative traits [e.g., Amos, 1994; Almasy et al., 1997; Blangero and Almasy, 1997; Almasy and Blangero, 1998; Amos et al., 2001; de Andrade et al., 1997, 2002], β corresponds to the variance of a random effect or the covariance matrix of several random effects, while γ corresponds to the fixed effects and possibly the covariance matrix of some other random effects; and we are interested in testing $H_0 : \beta = 0$ against $H_1 : \beta > 0$ in the former case, or $H_0 : \beta = 0$ against $H_1 : \beta$ is a positive semidefinite matrix in the latter case.

PARAMETRIC STATISTICS

We first consider the likelihood-based linkage analysis. Write $\theta = (\beta, \gamma)$. Let $l(\theta; d)$ be the natural logarithm of the likelihood function for θ calculated at a given genome position d. Then the likelihood ratio test statistic at position d takes the form $LR(d) = 2\{l(\widehat{\theta}; d) - l(0, \widetilde{\gamma}; d)\}$, where $\widehat{\theta}$ is the (unrestricted) maximum likelihood estimator of θ, and $\widetilde{\gamma}$ is the restricted maximum likelihood estimator of γ under $H_0 : \beta = 0$. Clearly, $LR(d) = (2 \log_e 10)\mathrm{LOD}(d)$, where $\mathrm{LOD}(d)$ is the classical LOD score [Morton, 1955] calculated at location d. It is assumed throughout this article that the number of pedigrees n is large. If we have two-sided alternatives, i.e., $H_1 : \beta \neq 0$, the statistic $LR(d)$ is approximately χ^2_p, where p is the dimension of β [Cox and Hinkley, 1974, §9.3]; otherwise, $LR(d)$ is approximately a weighted sum of chi-square random variables with different degrees of freedom [Self and Liang, 1987].

To assess the genomewide statistical significance, we need to evaluate the probability distribution of $LR(d)$ as a stochastic process indexed by d. For this purpose, it is more convenient to work with the score statistic for testing the same hypothesis. Another motivation for considering the score statistic is because of its close connection with nonparametric linkage statistics, as will be discussed later.

Let $U(d)$ be the score function for β at location d that is evaluated at $\beta = 0$ and $\gamma = \widetilde{\gamma}$. We show in the Appendix that approximately

$$U(d) = \sum_{i=1}^{n} U_i(d), \qquad (1)$$

where the $U_i(d)$ are given in Equation (A1) of the Appendix. Furthermore, the random vector $n^{-1/2}U(d)$ is approximately normal, with mean zero and with covariance matrix $\widehat{V}(d) = n^{-1}\sum \widehat{U}_i(d)\widehat{U}_i^T(d)$, where the $\widehat{U}_i(d)$ are given in Equation (A2) of the Appendix. The score statistic for testing $H_0 : \beta = 0$ against $H_1 : \beta \in C$ at location d takes the form

$$\begin{aligned} W(d) =& n^{-1}U^T(d)\widehat{V}^{-1}(d)U(d) \\ &- \min_{b \in C}[\{n^{-1/2}U(d) - \widehat{R}(d)b\}^T \widehat{V}^{-1}(d) \\ &\times \{n^{-1/2}U(d) - \widehat{R}(d)b\}] \end{aligned} \qquad (2)$$

where $\widehat{R}(d)$ is given in Equation (A3) of the Appendix. As indicated in the Appendix, the score statistic $W(d)$ is equivalent to the likelihood ratio statistic $LR(d)$ in large samples.

The evaluation of $W(d)$ in (2) involves the minimization of the quadratic form $\{n^{-1/2}U(d) - \widehat{R}(d)b\}^T\widehat{V}^{-1}(d)\{n^{-1/2}U(d) - \widehat{R}(d)b\}$, subject to the constraint $b \in C$. This is a simple minimization problem, for which special algorithms are available [see Wollan and Dykstra, 1987]. In most cases, the minimization is trivial. For two-sided alternatives, the minimum is always 0, so that $W(d)$ reduces to the familiar form $n^{-1}U^T(d)\widehat{V}^{-1}(d)U(d)$. For one-sided alternatives with a scalar parameter in the form of $H_1 : \beta > 0$, as in the case of a single variance component [Amos, 1994; Blangero and Almasy, 1997; Almasy and Blangero, 1998], we have $W(d) = n^{-1}U^2(d)/\widehat{V}(d)$ if $U(d) > 0$ and $W(d) = 0$ if $U(d) \leq 0$. For one-sided alternatives involving multiple parameters, such as multiple variance components [Almasy et al., 1997; Blangero and Almasy, 1997; de Andrade et al., 1997, 2002; Amos et al., 2001], it may be necessary to actually evaluate the second term of (2).

NONPARAMETRIC STATISTICS

Some linkage statistics, such as those of Haseman and Elston [1972], Kruglyak and Lander [1995a,b], and Kruglyak et al. [1996], are not derived from the likelihood. Those statistics, however, can all be written in the form of Equation (1), where U_i involves the data from the ith pedigree only. Thus, the general theory established in the Appendix implies that the nonparametric statistics can be treated in the same manner as the score statistic.

For the Haseman-Elston regression, $U(d) = \sum \widehat{U}_i(d)$, where $\widehat{U}_i(d) = (Y_i - \overline{Y})\{X_i(d) - \overline{X}(d)\}$. In this expression, Y_i is a function of the trait values of the ith sib pair, $X_i(d)$ is a function of the distribution of the number of alleles shared identical by decent (IBD) for the ith sib pair at location d, $\overline{Y} = n^{-1}\sum Y_i$, and $\overline{X}(d) = n^{-1}\sum X_i(d)$. For this type of regression, $\widehat{R}(d) = n^{-1} \times \sum\{X_i(d) - \overline{X}(d)\}\{X_i(d) - \overline{X}(d)\}^T$. In the traditional Haseman-Elston regression, Y_i is the squared difference between the trait values of the ith sib pair, and the covariate is the estimated proportion of the alleles IBD. Elston et al. (2000) recommended using the mean-corrected cross-product of the trait values rather than the squared difference, and including the estimated probability that the sib pair shares 2 alleles IBD as a second covariate so as to study additive and dominant genetic variances. In the latter case, β has two components with $\beta \geq 0$, and the tests should be one-sided in this direction.

Kruglyak and Lander [1995a] proposed the use of the EM algorithm to perform the Haseman-Elston regression on the actual IBD distribution; their analysis is based on the likelihood and is thus covered by the above results for the likelihood-based analysis. Olson and Wijsman [1993] and others generalized the Haseman-Elston regression to accommodate all relative pairs. Those statistics can also be written in the form of Equation (1), where U_i represents contributions from all relative pairs of the ith pedigree.

The nonparametric statistics advocated by Kruglyak et al. [1996] take the form

$$U(d) = \sum_{i=1}^{n} w_i\{S_i(d) - \mu_i(d)\}/\widehat{\sigma}_i(d)$$

where $S_i(d)$ is a scalar function of the IBD allele sharing for the ith pedigree at location d, which has a higher expected value under linkage than under no linkage, $\mu_i(d)$ and $\widehat{\sigma}_i(d)$ are the expected value and (estimated) standard derivation of $S_i(d)$ under no linkage, and the w_i are weighting factors such that $\sum w_i^2 = 1$. In this case, $W(d) = U^2(d)$ if $U(d) > 0$ and $W(d) = 0$ if $U(d) \leq 0$. When the estimates $\widehat{\sigma}_i(d)$ are not sufficiently accurate, we set $W(d) = U^2(d)/\widehat{V}(d)$ if $U(d) > 0$, and $W(d) = 0$ if $U(d) \leq 0$, where $\widehat{V}(d) = \sum w_i^2\{S_i(d) - \mu_i(d)\}^2/\widehat{\sigma}_i^2(d)$. Kong and Cox [1997] proposed a one-parameter model to allow missing data. Their method is likelihood-based and is thus covered by our results for parametric statistics.

GENOMEWIDE STATISTICAL SIGNIFICANCE

We now demonstrate how to assess genomewide statistical significance for an arbitrary linkage statistic. This assessment requires evaluation of the probability distribution of the maximum deviation statistic $\max_d W(d)$. To this end, we study the large-sample distribution of $U(d)$ regarded as a stochastic process in d. It is shown in the Appendix that $U(d)$ is approximately a zero-mean Gaussian process. In general, the probability distribution of $\max_d W(d)$ is not analytically tractable. We propose a Monte Carlo approach similar to that of Lin et al. [1993] to approximate the distribution of $\max_d W(d)$.

Define

$$\widehat{U}(d) = \sum_{i=1}^{n} \widehat{U}_i(d) G_i$$

where G_i $(i = 1, \ldots, n)$ are independent standard normal random variables that are independent of the data. Let

$$\begin{aligned}\widehat{W}(d) =& n^{-1}\widehat{U}^T(d)\widehat{V}^{-1}(d)\widehat{U}(d) \\ &- \min_{b \in C}[\{n^{-1/2}\widehat{U}(d) - \widehat{R}(d)b\}^T\widehat{V}^{-1}(d) \\ &\times\{n^{-1/2}\widehat{U}(d) - \widehat{R}(d)b\}].\end{aligned} \quad (3)$$

In (3), we regard $\widehat{V}(d)$, $\widehat{R}(d)$, and the $\widehat{U}_i(d)$ in $\widehat{U}(d)$ as fixed, and the G_i as random. To approximate the distribution of $\max_d W(d)$, we generate the normal random sample $(G_1, \cdots, G_n)$ a large number of times, say 1,000 or 10,000 times; for each sample, we calculate $\widehat{W}(d)$ and compare $\max_d \widehat{W}(d)$ with the observed value of $\max_d W(d)$. The proportion of the simulated $\max_d \widehat{W}(d)$ that are greater than or equal to the observed value of $\max_d W(d)$ is the genomewide p-value, and the $100(1-\alpha)$th percentile of the simulated $\max_d \widehat{W}(d)$ is the threshold value for the genomewide significance level of α.

Numerically, $\widehat{W}(d)$ is evaluated in the same manner as $W(d)$. For the likelihood-based analysis, one usually uses the likelihood ratio statistic $LR(d)$ rather than the score statistic $W(d)$, although the two statistics are asymptotically equivalent. In that case, the expression for $W(d)$ given in (2) provides some insights into the use of $\widehat{W}(d)$ given in (3) to approximate the distribution of $LR(d)$, although $W(d)$ itself needs not be directly evaluated. In other cases, both $W(d)$ and $\widehat{W}(d)$ need to be evaluated, and the connection between the test statistic and the simulated statistic is more transparent.

Unlike permutation tests and other simulation methods, the proposed Monte Carlo procedure involves simulation of normal random variables rather than the genotype or phenotype data, and does not require repeated analysis of simulated data sets. The quantities involving the observed data, i.e., $\widehat{V}(d)$, $\widehat{R}(d)$, and the $\widehat{U}_i(d)$, are only calculated once, and the evaluation of the empirical distribution of $\widehat{W}(d)$ given these quantities is trivial. Thus, the proposed approach is much less time-consuming than permutation tests and other simulation methods. More important, it applies to any linkage statistics, regardless of the inheritance patterns and pedigree structures.

RESULTS

SIMULATION STUDIES

We carried out extensive simulation studies to assess the performance of the proposed methods in practical situations and to demonstrate their advantages over the analytical methods. For the first series of studies, we considered one chromosome with a total length of 100 cM. We chose markers that are evenly spaced with a distance of 1, 5, 10, or 20 cM. We also chose unevenly spaced markers located either at 0, 20, 40, 47, 50, 55, 60, 65, 70, 80, and 100 cM or at 0, 5, 10, 20, 40, 60, 70, 80, 90, 95, and 100 cM, the average marker distance being 10 cM in both cases. In order to apply the analytical formulas, we generated parental alleles with infinitely many alleles so that the genotypes of the markers are fully informative. No matter where the markers were located, we placed the major QTL at 45 cM.

We considered 200 sib pairs, and generated two quantitative traits as follows. Under the null hypothesis of no linkage, the traits are multivariate normal, with means 0 and variances 1; for the first trait, the correlation between the trait values of the sib pair is 0.3; the two traits of the same individual have a correlation of 0.2. Under the alternative hypothesis, genetic effects are added to the normal random errors generated under the null hypothesis; the two traits are controlled by the same QTL with common genetic effects. We considered three sets of QTL effects specified by the additive genetic variance σ_g^2 and dominant genetic variance σ_d^2: (i) $\sigma_a^2 = 2$, $\sigma_d^2 = 0$; (ii) $\sigma_a^2 = 0$, $\sigma_d^2 = 2$; and (iii) $\sigma_a^2 = \sigma_d^2 = 1$.

We implemented the Haseman-Elston regression with both the univariate and bivariate traits. Following the recommendation of Elston et al. [2000], we used the mean-corrected cross-product, rather than the squared difference, of the trait values of the sib pair. Note that such a response variable has a very skewed distribution. The univariate analysis is performed on the first trait. For the bivariate analysis, $U(d)$ has two components, one being the (new) Haseman-Elston statistic for the first trait, and one being that of the second trait.

For each combination of the genetic map and QTL effects, we simulated 10,000 data sets. For each simulated data set, we evaluated the statistic $W(d)$ at a 1-cM increment and calculated its maximum over all testing positions. The targeted genomewide significance level or type 1 error α was set at 0.01 or 0.05. The thresholds based on the proposed Monte Carlo approach were determined by generating the normal random samples 10,000 times. For the univariate Haseman-Elston regression, we calculated the analytical thresholds according to the dense-marker formula of Lander and Kruglyak [1995] and the sparse-marker approximations of Dupuis and Siegmund [1999].

TABLE I. Thresholds for univariate and bivariate Haseman-Elston regression

		Empirical[b]		Proposed[c]		Dense[d]		Sparse[e]	
Analysis	Marker space[a]	α=0.05	0.01	α=0.05	0.01	α=0.05	0.01	α=0.05	0.01
Univariate	1	8.05	11.42	7.73 (0.25)	11.14 (0.30)	8.99	12.58	8.06	11.52
	5	7.19	10.41	6.97 (0.21)	10.29 (0.28)	8.99	12.58	6.98	10.23
	10	6.78	10.32	6.50 (0.20)	9.77 (0.26)	8.99	12.58	6.29	9.40
	20	6.04	9.30	5.96 (0.18)	9.18 (0.26)	8.99	12.58	5.48	8.39
	U1	6.51	9.87	6.37 (0.20)	9.63 (0.27)	8.99	12.58	6.29	9.40
	U2	6.63	9.94	6.42 (0.19)	9.68 (0.26)	8.99	12.58	6.29	9.40
Bivariate	1	10.09	13.35	10.03 (0.21)	13.64 (0.28)				
	5	9.20	12.41	9.17 (0.19)	12.71 (0.26)				
	10	8.64	12.12	8.64 (0.18)	12.13 (0.26)				
	20	8.03	11.50	8.04 (0.17)	11.51 (0.25)				
	U1	8.65	11.72	8.50 (0.18)	12.00 (0.26)				
	U2	8.56	11.92	8.55 (0.18)	12.04 (0.26)				

[a] Markers are evenly spaced with a distance of 1, 5, 10, or 20 cM, or unevenly spaced at $\mathrm{U1} = \{0, 20, 40, 47, 50, 55, 60, 65, 70, 80, 100\}$ cM or $\mathrm{U2} = \{0, 5, 10, 20, 40, 60, 70, 80, 90, 95, 100\}$ cM.

[b] The $(1 - \alpha)$100th percentiles of test statistics based on 10,000 simulated data sets.

[c] Means of proposed thresholds over 10,000 simulated data sets, with standard derivations in parentheses.

[d] Dense-marker thresholds of Lander and Kruglyak [1995].

[e] Sparse-marker thresholds of Dupuis and Siegmund [1999].

TABLE II. Empirical type 1 errors and powers for Haseman-Elston regression

			Proposed methods				Analytical methods			
			Univariate[b]		Bivariate[c]		Dense[d]		Sparse[e]	
σ_a^2	σ_d^2	Marker space[a]	α=0.05	0.01	α=0.05	0.01	α=0.05	0.01	α=0.05	0.01
0	0	1	0.059	0.012	0.051	0.009	0.031	0.006	0.050	0.010
		5	0.056	0.012	0.051	0.009	0.023	0.004	0.056	0.011
		10	0.058	0.013	0.050	0.009	0.017	0.004	0.064	0.015
		20	0.052	0.011	0.050	0.010	0.013	0.002	0.067	0.017
		U1	0.055	0.011	0.055	0.009	0.016	0.003	0.056	0.013
		U2	0.056	0.012	0.047	0.010	0.017	0.003	0.059	0.014
2	0	1	0.756	0.513	0.948	0.816	0.665	0.417	0.734	0.486
		5	0.748	0.512	0.952	0.832	0.603	0.366	0.745	0.513
		10	0.700	0.452	0.912	0.740	0.506	0.277	0.716	0.477
		20	0.665	0.411	0.883	0.691	0.425	0.218	0.707	0.470
		U1	0.761	0.515	0.943	0.813	0.562	0.323	0.767	0.531
		U2	0.637	0.382	0.863	0.660	0.427	0.222	0.648	0.400
0	2	1	0.544	0.300	0.738	0.478	0.442	0.219	0.517	0.275
		5	0.538	0.289	0.750	0.484	0.374	0.175	0.537	0.289
		10	0.498	0.250	0.691	0.423	0.298	0.126	0.516	0.271
		20	0.467	0.225	0.659	0.379	0.234	0.095	0.511	0.270
		U1	0.547	0.298	0.746	0.485	0.337	0.150	0.552	0.310
		U2	0.447	0.213	0.627	0.355	0.250	0.101	0.457	0.226
1	1	1	0.684	0.433	0.893	0.706	0.589	0.335	0.658	0.401
		5	0.679	0.429	0.902	0.722	0.525	0.291	0.678	0.432
		10	0.630	0.371	0.848	0.624	0.424	0.217	0.647	0.396
		20	0.586	0.334	0.814	0.576	0.348	0.167	0.630	0.388
		U1	0.679	0.428	0.891	0.707	0.470	0.250	0.686	0.443
		U2	0.562	0.315	0.785	0.537	0.356	0.169	0.572	0.332

[a] Markers are evenly spaced with a distance of 1, 5, 10, or 20 cM, or unevenly spaced at $\mathrm{U1} = \{0, 20, 40, 47, 50, 55, 60, 65, 70, 80, 100\}$ cM or $\mathrm{U2} = \{0, 5, 10, 20, 40, 60, 70, 80, 90, 95, 100\}$ cM.

[b] Univariate Haseman-Elston regression based on the proposed thresholds.

[c] Bivariate Haseman-Elston regression based on the proposed thresholds.

[d] Univariate Haseman-Elston regression based on dense-marker thresholds of Lander and Kruglyak [1995].

[e] Univariate Haseman-Elston regression based on sparse-marker thresholds of Dupuis and Siegmund [1999].

When the markers are unevenly spaced, the average marker distance is used in the sparse-marker approximations. No analytical thresholds are available for the bivariate Haseman-Elston regression.

Table I displays the proposed and analytical thresholds in the simulation studies. It also shows the empirical thresholds based on the 10,000 simulated data sets, which can be regarded as the true thresholds. The empirical thresholds provide a benchmark in the simulation studies, but would be unknown in real data analysis. Because they are data-dependent, the proposed thresholds vary from data set to data set. As shown in Table I, the mean values of the proposed thresholds match reasonably well with the empirical values, the standard deviations being small. By contrast, the thresholds based on the dense-marker theory are much larger than the empirical values, while the thresholds based on the sparse-marker approximations tend to be too small when the markers are sparse.

Table II reports the empirical type 1 errors and powers of the linkage tests, based on the proposed and analytical thresholds. The tests based on the proposed thresholds maintain type 1 errors near the targeted levels under all genetic maps, both for the univariate and bivariate analyses. The bivariate analysis offers substantial power advantages over the univariate analysis. The analytical thresholds based on the dense-marker theory yield overly conservative tests and consequently incur loss of power, even when the markers are as dense as 1 cM apart. The type 1 errors under the sparse-marker approximations tend to be higher than the targeted levels when the markers are sparse, especially for $\alpha = 0.01$.

To assess the accuracy of the asymptotic approximations in genome scans, we conducted a series of studies with the 22 human autosomal chromosomes. The physical lengths of the 22 chromosomes are approximately 2×240, 200, 2×180, 2×160, 2×140, 3×120, 3×100, 4×80, and 3×60 Mb. The physical lengths were converted to the map distances by assuming that 1 cM corresponds to 1 Mb. We considered sib pairs, and generated a quantitative trait that is standard normal with a between-sibs correlation of 0.3. We chose equally spaced markers with intermarker distance of 1, 5, 10, or 20 cM, and performed the Haseman-Elston regression at a 1-cM increment. For $\alpha = 0.05$ and $n = 200$, the empirical type 1 errors based on 10,000 simulated data sets are 0.068, 0.067, 0.061, and 0.060 for marker distances of 1, 5, 10, and 20 cM, respectively. For $n = 400$, the corresponding estimates are 0.059, 0.057, 0.052, and 0.058. Since about 3,000 tests are performed across the genome, the thresholds pertain to the extreme tail of the distribution, for which the normal approximation tends to be poor. Thus, it is remarkable that the proposed thresholds are only slightly liberal.

GAW10 NUCLEAR FAMILY DATA

To illustrate the proposed approach in a complex setting, we consider the nuclear family data from the Tenth Genetic Analysis Workshop (GAW10). The data were generated by computer simulation for a set of 239 nuclear families with a total of 1,164 individuals. The number of sibs in a family ranges from 2–6, with an average of 2.87. The simulation, described in detail by MacCluer et al. [1997], included five quantitative traits (Q1–Q5) controlled by six underlying major genes (MG1–MG6) plus additional polygenic determinants. We confine our attention to Q4 and Q5: Q4 is influenced by MG4, MG5, and MG6, which are located on chromosomes 8, 9, and 10, respectively, while Q5 is influenced by MG4 and MG5. The residual environmental correlation between Q4 and Q5 is approximately 0.4. For each individual, there are 367 highly polymorphic markers on 10 chromosomes, with 24–50 markers per chromosome and with an average marker space of 2.03 cM. The total length of the genome is 726 cM. The markers have 4 to 15 alleles (mean=6.7), with a mean heterozygosity of 0.77. Phenotypes are available for 1,000 individuals, who are living. There are 200 replicates of the simulated data.

The simulated data have complexities similar to real data but with known answers, and thus are useful in testing and comparing statistical methods. Wijsman and Amos [1997] summarized the different approaches taken by the GAW10 participants. The simulated data have subsequently been used by many authors to validate new linkage methods. There has been no formal assessment of genomewide statistical significance for these data. Most authors simply regarded a LOD score of above 3 or 1.8, say, as evidence for "significant" or "suggestive" evidence of linkage [Williams and Blangero, 1999; Wang et al., 2001].

We illustrate the proposed approach by applying the univariate and bivariate Haseman-Elston regression methods to traits Q4 and Q5. We calculate the average IBD allele sharing at a 1-cM increment by using the multipoint probabilities that two sibs share 0, 1, or 2 alleles IBD obtained from the Mapmaker/Sibs program [Kruglyak and Lander, 1995a]. All possible sib pairs in each family are constructed and used in the analysis. The contribution from a family to the statistic pertains to the sum of the mean-corrected cross-products of all the sib pairs in that family. For the bivariate Haseman-Elston regression, the statistic has two components corresponding to Q4 and Q5. We divide all the test statistics by $2 \log_e 10$ to convert them to conventional LOD scores. The means of the proposed thresholds for $\alpha = 0.05$ over the 200 replicates turn out to be 2.45 with a standard deviation of 0.027 for the univariate analysis of Q4, 2.45 with a standard deviation of 0.03 for the univariate analysis of Q5, and 3.04 with a standard deviation of 0.027 for the joint analysis of Q4 and Q5. The corresponding mean threshold values for $\alpha = 0.01$ are 3.17, 3.18, and 3.81, respectively, with standard deviations of 0.049, 0.049, and 0.051.

Figure 1 displays the mean LOD scores over the 200 replicates at all genome positions. The mean LOD scores have three peaks around the three QTLs, although the values of the peaks are below the thresholds. In our analysis, we group the consecutive loci at which the LOD scores exceed the threshold and refer to them as one significant region. At $\alpha = 0.05$, the univariate analysis of Q4 identifies 38 significant regions in 37 replicates, while the analysis of Q5 identifies 25 significant regions in 24 replicates. In the analysis of Q4, most of the regions identified are close to MG4 located on chromosome 8. Most of the regions

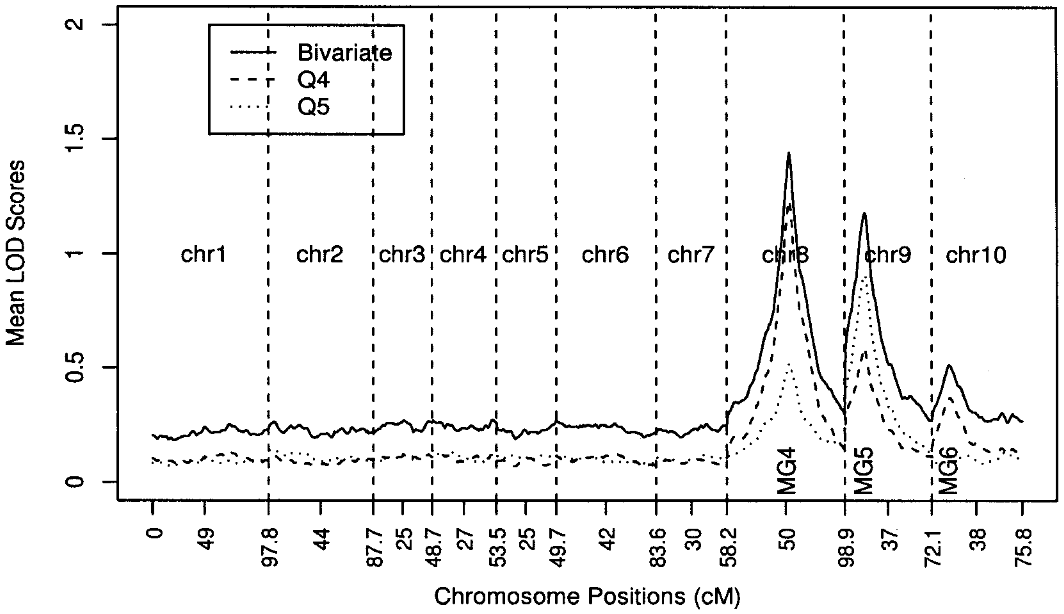

Fig. 1. Mean LOD scores for univariate and bivariate Haseman-Elston regression analyses of quantitative traits Q4 and Q5: mean LOD scores over 200 replicates of GAW10 data are plotted across 10 chromosomes, which are ordered from left to right and separated by dashed vertical lines. MG4, MG5, and MG6 are three major QTLs.

identified by the analysis of Q5 are close to MG5, located on chromosome 9. These results are consistent with the underlying genetic model, as MG4 is the QTL with the highest genetic contribution to Q4, while MG5 has the highest contribution to Q5.

The joint analysis of Q4 and Q5 identifies 36 significant regions. The reason that the joint analysis is not more powerful than the univariate analysis of Q4 is that, unlike our simulation studies, the genetic effects of MG4, MG5, and MG6 on Q4 and Q5 are very different. Nevertheless, the joint analysis preserves an overall type 1 error. The power of the univariate analysis would be reduced substantially if the Bonferroni correction were used to adjust for the fact that two traits are examined in the same study.

To compare the proposed and analytical methods, we calculate the dense-map threshold for the targeted significance level of 0.05, which turns out to be 3.19. If this threshold is used for assessing the genomewide significance of the test statistic in the univariate analysis of Q4, we will identify only 7, instead of 38, significant regions. Thus, the use of the analytical thresholds results in considerable reduction of power.

Figure 2 plots the LOD scores for replicates 22 and 41 from the univariate analysis of Q4 on chromosomes 8, 9, and 10, along with the proposed and analytical thresholds for $\alpha = 0.05$. The proposed method successfully identifies one significant region around MG4 on chromosome 8 in replicate 22, and two significant regions around MG4 and MG5 in replicate 41. The analytical method identifies the same regions, but almost misses MG5 in replicate 41. The analytical thresholds are considerably higher than the proposed ones. The genomewide p-values for the proposed and analytical methods are, respectively, 0.002 and 0.006 in replicate 22, and 0.002 and 0.005 in replicate 41.

Figure 3 shows the results for the univariate analysis of Q4 in replicates 5 and 62. In these two cases, the proposed method successfully identifies MG4, whereas the analytical method fails to. The analytical thresholds are again much higher than the proposed ones. The genomewide p-values for the proposed and analytical methods are, respectively, 0.019 and 0.054 in replicate 5, and 0.021 and 0.056 in replicate 62.

The above results indicate that the Haseman-Elston regression has relatively low power in detecting genetic linkage for the GAW10 data. Williams and Blangero [1999] and Korczak and Goldstein [1997] reached similar conclusions from their analysis of Q1. When the power is low, it is critical to assess the genomewide significance in an accurate manner so that important signals would not be missed due to the use of overly conservative tests.

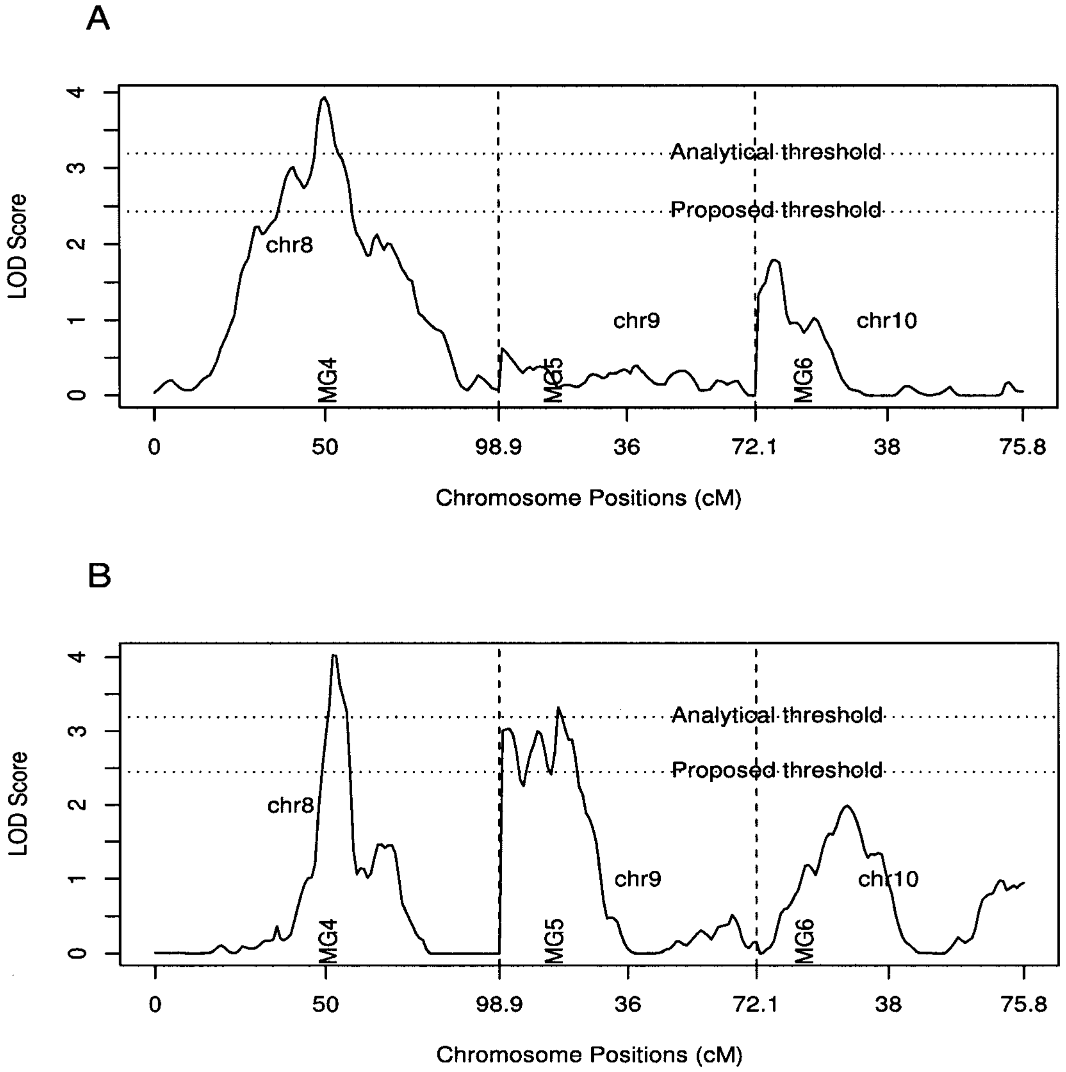

Fig. 2. LOD score across chromosomes 8, 9, and 10 for univariate Haseman-Elston regression analysis of quantitative Q4 in replicates 22 (A) and 41 (B) of GAW10 data. Two dotted horizontal lines pertain to analytical and proposed thresholds at targeted 5% genomewide significance level. MG4, MG5, and MG6 are major QTLs for Q4.

DISCUSSION

The proposed approach has a distinct advantage over the existing methods in that it is completely general, being applicable to any type of linkage study and any data structure, provided that the number of pedigrees is reasonably large. It is more accurate than the analytical solutions because it takes into consideration the actual data structure rather than relying on idealized conditions that may differ substantially from the actual conditions. It is less time-consuming than the existing simulation-based methods, because only normal random variables are simulated; for each replicate of the GAW10 data, calculating the thresholds based on 10,000 normal samples takes half a minute on an IBM BladeCenter HS20 machine. Most important, the proposed approach can deal with many study designs, data structures, and genetic models to which the existing methods are not applicable.

As pointed out by Lander and Kruglyak [1995], it is difficult to derive analytical thresholds if investigators conduct several analyses by trying out multiple diagnostic schemes for defining affectation status, multiple models of inheritance, or multiple traits. The current practice is to regard each of the *K* analyses as statistically independent and to apply the Bonferroni correction. This prescription is overly conservative, since the

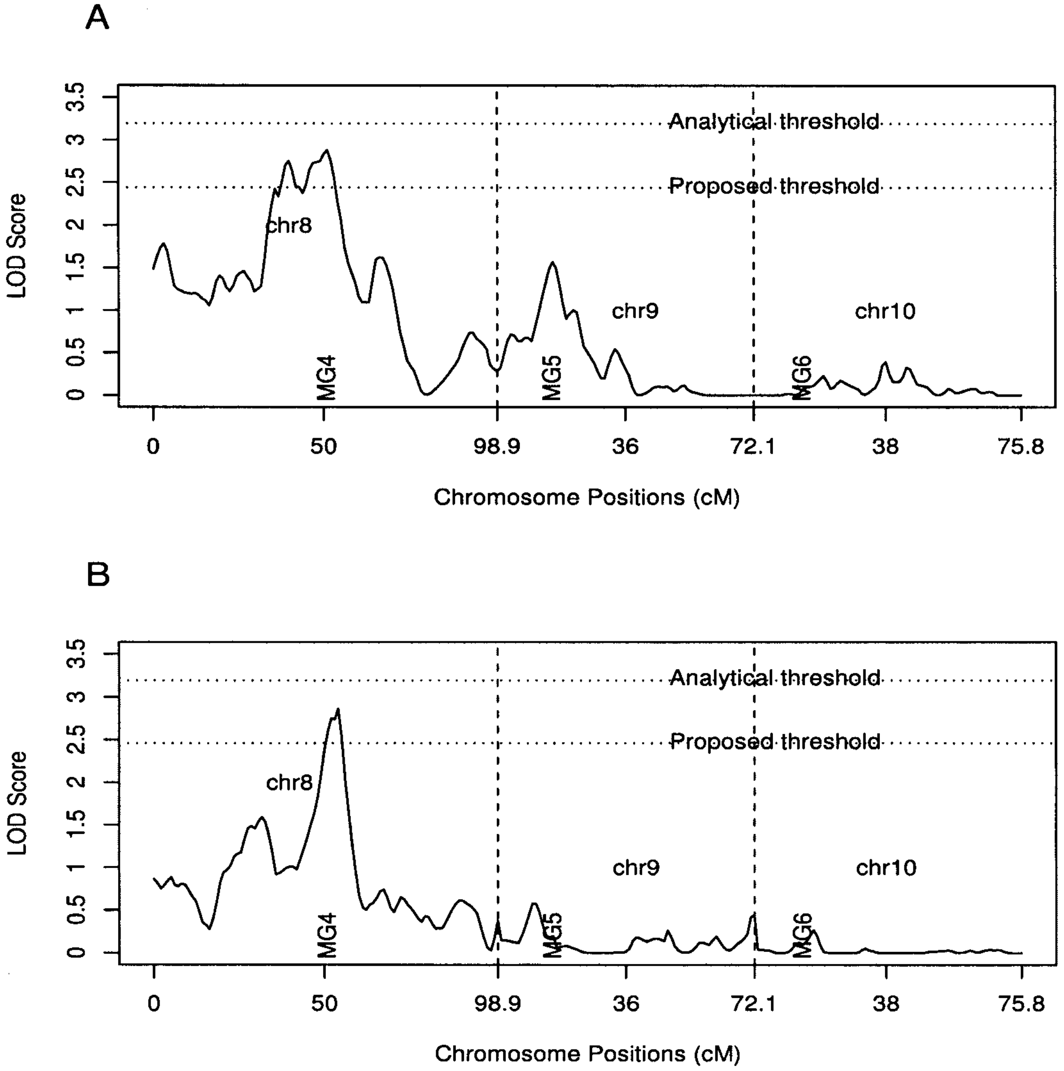

Fig. 3. LOD score across chromosomes 8, 9, and 10 for univariate Haseman-Elston regression analysis of quantitative trait Q4 in replicates 5 (A) and 62 (B) of GAW10 data. Two dotted horizontal lines pertain to analytical and proposed thresholds at targeted 5% genomewide significance level. MG4, MG5, and MG6 are major QTLs for Q4.

analyses tend to be correlated. It is straightforward to apply our approach to this setting. Each of the K statistics can be written in the form of Equation (1). We can stack up these K statistics and form an overall score-type statistic, as we did for the multivariate Haseman-Elston regression, or consider $\max_k \max_d W_k(d)$, where $W_k(d)$ is the score-type test statistic for the kth analysis in the form of (2). We can then use the proposed approach to determine the thresholds.

Our approach can be extended to the situations of "dependent" families (e.g., genotyped individuals appearing in multiple extended families). Suppose that there are n independent sets of families and, for $i = 1, \dots, n$, the ith set contains n_i dependent families. The test statistic takes the form $U(d) = \sum_{i=1}^{n} \sum_{j=1}^{n_i} U_{ji}(d)$, where U_{ji} is the contribution from the jth family of the ith set. We simulate from the distribution of $\widehat{U}(d) = \sum_{i=1}^{n} \sum_{j=1}^{n_j} \widehat{U}_{ji}(d)G_i$, where $\widehat{U}_{ji}$ is the empirical counterpart of U_{ji}, and $(G_1, \dots, G_n)$ are independent standard normal.

A potential drawback of empirical methods, including the proposed methods and all simulation-based methods, is that the thresholds depend on the observed data, and are thus unknown in the design stage. One strategy is to use the analytical thresholds based on the sparse-marker approximations in the design stage and use the empirical thresholds in the analysis stage; when analytical thresholds are not available, simulation methods can be used to estimate thresholds in the design stage by mimicking the actual study. This is a common practice in other related contexts, such as sequential clinical trials.

It is important to distinguish between the number of markers and the number of testing positions. The dense-marker theory assumes that the markers are infinitely dense and the linkage tests are performed at all markers. The current genotyping technology does not permit infinitely dense marker data for the entire genome, so that the thresholds based on the dense-marker theory tend to be overly conservative. The analytical solutions for the sparse-marker approximations assume that the tests are performed at the marker locations only, although the actual tests are usually performed between markers and thus have more complicated distributions. As a result, the thresholds based on the sparse-marker approximations tend to be too liberal. By contrast, the proposed methods yield accurate thresholds for the actual tests, whether they are performed at the markers or between the markers, or both. The true thresholds depend on the marker data as well as the testing positions, and will not remain the same if one changes the marker density while keeping the testing positions fixed or vice versa.

The assessment of statistical significance in linkage analysis appears to be a contentious issue [Lander and Kruglyak, 1995, 1996; Curtis, 1996; Witte et al., 1996; Sawcer et al., 1997; Kruglyak and Daly, 1998]. The principle underlying the proposed approach is that the effects of multiple testing should be adjusted for, and the adjustment depends on what data are eventually observed and how the tests are actually performed. If the study involves only a genome scan with markers every 10 cM, then the proposed thresholds will reflect such a sparse map. If one employs a hierarchical search, in which one performs a genome scan with a sparse map and then follows up interesting regions with a denser map, then the proposed thresholds will reflect the fact that the map is dense among the regions chosen for follow-up and sparse elsewhere. As mentioned above, the testing positions are usually dense, whether the markers are dense or not. The proposed methods properly account for the observed marker data and actual testing positions, and thus produce accurate thresholds and proper significance levels.

We have written some programs to implement the proposed methods. Since there already exist many excellent programs for linkage analysis, the most efficient way of implementing the proposed methods is to incorporate our Monte Carlo procedure into the existing programs. In principle, this is a simple task, since all the ingredients required for the Monte Carlo procedure are available in the existing linkage analysis programs. We are currently working on the interface between GENHUNTER [Kruglyak et al., 1996] and our Monte Carlo procedure, and will make the resulting programs available to the public in the near future.

ACKNOWLEDGMENTS

The authors are grateful to the reviwers for their comments, and to Jean W. MacCluer for providing the nuclear family data from GAW10, which was supported in part by grant GM31575 from the National Institutes of Health.

REFERENCES

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1121.

Almasy L, Dyer TD, Blangero J. 1997. Bivariate quantitative trait linkage analysis: pleiotropy versus coincident linkages. Genet Epidemiol 14:953–958.

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543.

Amos CI, de Andrade M, Zhu D. 2001. Comparison of multivariate tests for genetic linkage. Hum Hered 51:133–144.

Blangero J, Almasy L. 1997. Multipoint oligogenic linkage analysis of quantitative traits. Genet Epidemiol 14:959–964.

Churchill GA, Doerge RW. 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Cox DR, Hinkley DV. 1974. Theoretical statistics. New York: Chapman and Hall.

Curtis D. 1996. Genetic dissection of complex traits. Nat Genet 12:356–357.

Davis S, Schroeder M, Goldin LR, Weeks DE. 1996. Nonparametric simulation-based statistics for detecting linkage in general pedigrees. Am J Hum Genet 58:867–880.

de Andrade M, Thiel TJ, Yu LP, Amos CI. 1997. Assessing linkage on chromosome 5 using components of variance approach: univariate versus multivariate. Genet Epidemiol 14:773–778.

de Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos CI. 2002. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol 22:221–232.

Doerge RW, Churchill GA. 1996. Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285–294.

Dupuis J, Siegmund D. 1999. Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373–386.

Elston RC, Buxbaum S, Jacobs KB, Olston JM. 2000. Haseman and Elston revisited. Genet Epidemiol 19:1–17.

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 53:234–251.

Guerra R, Wan Y, Jia A, Amos CI, Cohen JC. 1999. Testing for linkage under robust genetic models. Hum Hered 49:146–153.

Haseman JK, Elston RC. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behavior Genetics 2:3–19.

Jansen RC, Stam P. 1994. High-resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447–1455.

Kao CH, Zeng ZB. 1997. General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics 53:653–665.

Kong A, Cox NJ. 1997. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–1188.

Korczak JF, Goldstein AM. 1997. Sib-pair linkage analyses of nuclear family data: quantitative versus dichotomous disease classification. Genet Epidemiol 14:827–832.

Kruglyak L, Daly MJ. 1998. Linkage threshold for two-stage genome scans. Am J Hum Genet 62:994–996.

Kruglyak L, Lander ES. 1995a. Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am J Hum Genet 57:439–454.

Kruglyak L, Lander ES. 1995b. A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421–1428.

Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–1363.

Lander ES, Botstein D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199.

Lander E, Kruglyak L. 1995. Genetic dissection of complex traits-guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247.

Lander E, Kruglyak L. 1996. Genetic dissection of complex traits. Nat Genet 12:355–356.

Lander ES, Schork NJ. 1994. Genetic dissection of complex traits. Science 265:2037–2048.

Lin DY, Wei LJ, Ying Z. 1993. Checking the Cox model with cumulative sums of Martingale-based residuals. Biometrika 80:557–572.

MacCluer JW, Blangero J, Dyer TD, Speer MC. 1997. GAW10: simulated family data for a common oligogenic disease with quantitative risk factors. Genet Epidemiol 14:737–742.

Morton NE. 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7:277–318.

Olson JM, Wijsman EM. 1993. Linkage between quantitative trait and marker locus methods using all relative pairs. Genet Epidemiol 10:87–102.

Ott J. 1989. Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 89:4175–4178.

Sawcer S, Jones HB, Judge D, Visser F, Compston A, Goodfellow PN, Clayton D. 1997. Empirical genomewide significance levels established by whole genome simulations. Genet Epidemiol 14:223–229.

Self SG, Liang KY. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610.

Silvapulle MJ, Silvapulle P. 1995. A score test against one-sided alternatives. J Am Stat Assoc 90:342–349.

Song KK, Weeks DE, Sobel E, Feingold E. 2004. Efficient simulation of *P* values for linkage analysis. Genet Epidemiol 26:88–96.

van der Vaart AW. 1998. Asymptotic statistics. Cambridge: Cambridge University Press.

Wan Y, Cohen J, Guerra R. 1997. A permutation test for the robust sib-pair linkage method. Ann Hum Genet 61:79–87.

Wang D, Lin S, Cheng R, Gao X, Wright FA. 2001. Transformation of sib-pair values for the Haseman-Elston method. Am J Hum Genet 68:1238–1249.

Weeks DE, Ott J, Lathrop GM. 1990. SLINK: a general simulation program for linkage analysis. Am J Hum Genet [Suppl] 47:204.

Wijsman EM, Amos CI. 1997. Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 14:719–735.

Williams JT, Blangero J. 1999. Joint multipoint linkage analysis of multivariate qualitative and quantitative traits. I. Likelihood formulation and simulation results. Genet Epidemiol 16: 113–134.

Witte JS, Elston RC, Schork NJ. 1996. Genetic dissection of complex traits. Nat Genet 12:355–356.

Wollan PC, Dykstra RL. 1987. Minimizing linear inequality contrained Mahalanobis distances. Appl Stat 36:234–240.

Zeng ZB. 1994. Precision mapping of quantitative traits loci. Genetics 136:1457–1468.

Zhao H, Merikangas KR, Kidd KK. 1999. On a randomization procedure in linkage analysis. Am J Hum Genet 65: 1449–1456.

Zou F, Yandell BS, Fine JP. 2001. Statistical issues in the analysis of quantitative traits in combined crosses. Genetics 158: 1339–1346.

APPENDIX

DERIVATION OF THEORETICAL RESULTS

The vector-valued score function for θ calculated at location d takes the form $S(\theta;d) = \partial l(\theta;d)/\partial\theta$, where $l(\theta;d)$ is the log-likelihood function. The score function is one example of estimating functions. In some situations, such as the Haseman-Elston regression, we are interested in estimating functions that are not based on parametric likelihood. In the sequel, we let $S(\theta;d)$ denote an arbitrary unbiased estimating function for θ calculated at location d. In general, $S(\theta;d) = \sum_{i=1}^{n} S_i(\theta;d)$, where S_i involves the data from the ith pedigree only. Let $\widehat{A}(\theta;d) = -n^{-1}\partial S(\theta;d)/\partial\theta$, and let $A(d)$ denote the limit of $\widehat{A}(\theta;d)$, as n tends to infinity. We partition the vector $S(\theta;d)$ into $S_\beta(\theta;d)$ and $S_\gamma(\theta;d)$, and partition the vector $S_i(\theta;d)$ into $S_{\beta,i}(\theta;d)$, and $S_{\gamma,i}(\theta;d)$ to conform with the partition (β,γ) of θ. Accordingly, we partition the matrix $A(d)$ into $A_{\beta\beta}(d)$, $A_{\beta\gamma}(d)$, $A_{\gamma\beta}(d)$, and $A_{\gamma\gamma}(d)$, and partition the matrix $\widehat{A}(\theta;d)$ in the same manner.

To test the null hypothesis $H_0 : \beta = 0$, we consider the statistic $U(d) = S_\beta(0,\widetilde{\gamma};d)$, where $\widetilde{\gamma}$ is the solution to the equation $S_\gamma(0,\gamma;d) = 0$. It follows from the Taylor series expansions and the law of large numbers that, up to an asymptotically negligible term, $n^{-1/2}U(d) = n^{-1/2}\times$

$\sum_{i=1}^{n} U_i(d)$, where

$$U_i(d) = S_{\beta,i}(0,\gamma;d) - A_{\beta\gamma}(d)A_{\gamma\gamma}^{-1}(d)S_{\gamma,i}(0,\gamma;d). \qquad (A1)$$

Under the null hypothesis of no linkage, the $U_i(d)$ are independent zero-mean random vectors for any given d. Thus, it follows from the multivariate central limit theorem that $n^{-1/2}U(d)$, when regarded as a multidimensional stochastic process in d, is asymptotically a zero-mean Gaussian process whose covariance function between d_1 and d_2, denoted by $V(d_1, d_2)$, is the limit of $n^{-1}\sum_{i=1}^{n} U_i(d_1)U_i^T(d_2)$, provided that the process is tight. We can verify the tightness of $n^{-1/2}U(d)$ by noting that $U(d)$ is a function of the trait, which does not depend on d, and the genotype, which is a function of bounded variation in d, and then apply arguments given in Examples 19.11 and 19.20 of van der Vaart (1998). The replacements of the unknown parameters in (A1) with their sample estimators yield

$$\widehat{U}_i(d) = S_{\beta,i}(0,\widetilde{\gamma};d) - \widehat{A}_{\beta\gamma}(0,\widetilde{\gamma};d)\widehat{A}_{\gamma\gamma}^{-1}(0,\widetilde{\gamma};d)S_{\gamma,i}(0,\widetilde{\gamma};d). \qquad (A2)$$

By the law of large numbers and the consistency of $\widetilde{\gamma}$, the limiting covariance function $V(d_1, d_2)$ can be consistently estimated by $\widehat{V}(d_1, d_2) = n^{-1}\sum_{i=1}^{n} \widehat{U}_i(d_1)\widehat{U}_i^T(d_2)$. The foregoing results imply that, for a given d, the random vector $n^{-1/2}U(d)$ is asymptotically zero-mean normal, with a covariance matrix that is consistently estimated by $\widehat{V}(d) = \widehat{V}(d, d)$.

Define

$$\widehat{R}(d) = \widehat{A}_{\beta\beta}(0,\widetilde{\gamma};d) - \widehat{A}_{\beta\gamma}(0,\widetilde{\gamma};d)\widehat{A}_{\gamma\gamma}^{-1}(0,\widetilde{\gamma};d)\widehat{A}_{\gamma\beta}(0,\widetilde{\gamma};d) \qquad (A3)$$

and write $\widetilde{U}(d) = n^{-1/2}\widehat{R}^{-1}(d)U(d)$, and $\widetilde{V}(d) = \widehat{R}^{-1}(d)\widehat{V}(d)\widehat{R}^{-1}(d)$. The score-type statistic for testing $H_0 : \beta = 0$ against the general alternative $H_1 : \beta \in C$ at location d takes the form

$$W(d) = \widetilde{U}^T(d)\widetilde{V}^{-1}(d)\widetilde{U}(d) - \min_{b \in C}\{\widetilde{U}(d) - b\}^T\widetilde{V}^{-1}(d)\{\widetilde{U}(d) - b\}$$

[Silvapulle and Silvapulle, 1995], which can also be written as Equation (2) given in the text. If the estimating function $S(\theta)$ is the likelihood score function for θ, then the score statistic $W(d)$ is asymptotically equivalent to the likelihood ratio statistic $LR(d)$ [Cox and Hinkley, 1974, §9.3; Silvapulle and Silvapulle, 1995].

Recall that $\widehat{U}(d) = \sum_{i=1}^{n} \widehat{U}_i(d)G_i$, where the G_i are independent standard normal random variables that are independent of the observable data. Thus, conditional on the observable data, $\widehat{U}(d)$ is normal with mean 0 at each location d, and the covariance between $n^{-1/2}\widehat{U}(d_1)$ and $n^{-1/2}\widehat{U}(d_2)$ equals $\widehat{V}(d_1, d_2)$, which converges to $V(d_1, d_2)$. It follows that the conditional distribution of the process $n^{-1/2}\widehat{U}(d)$, given the observable data, converges to the distribution of the process $n^{-1/2}U(d)$. Consequently, the distribution of $W(d)$ can be approximated by that of $\widehat{W}(d)$.