*Am. J. Hum. Genet. 65:1449–1456, 1999*

# On a Randomization Procedure in Linkage Analysis

Hongyu Zhao,[1,2] Kathleen R. Merikangas,[1] and Kenneth K. Kidd[2]

[1]Department of Epidemiology and Public Health and [2]Department of Genetics, Yale University School of Medicine, New Haven, CT

## Summary

**Although much theoretical work has been undertaken to derive thresholds for statistical significance in genetic linkage studies, real data are often complicated by many factors, such as missing individuals or uninformative markers, which make the validity of these theoretical results questionable. Many simulation-based methods have been proposed in the literature to determine empirically the statistical significance of the observed test statistics. However, these methods either are not generally applicable to complex pedigree structures or are too time-consuming. In this article, we propose a computationally efficient simulation procedure that is applicable to arbitrary pedigree structures. This procedure can be combined with statistical tests, to assess the statistical significance for genetic linkage between a locus and a qualitative or quantitative trait. Furthermore, the genomewide significance level can be appropriately controlled when many linked markers are studied in a genomewide scan. Simulated data and a diabetes data set are analyzed to demonstrate the usefulness of this novel simulation method.**

## Introduction

In a typical genomewide scan, hundreds of markers are typed. Because of the lack of independence among these markers and the uncertainties in inferring the allele-sharing status at a given locus, the determination of the significance level of linkage has been an active area of research in human genetics. There appears to be no general agreement among statistical geneticists on the reportage of "significant" linkage findings (Lander and Kruglyak 1995; Curtis 1996; Witte et al. 1996).

Received September 2, 1998; accepted for publication July 20, 1999; electronically published October 14, 1999.

Address for correspondence and reprints: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, 60 College Street, Yale University School of Medicine, New Haven, CT 06520-8034. E-mail: hongyu.zhao@yale.edu


0002-9297/1999/6505-0029$02.00

The determination of significance levels varies among linkage programs available to the scientific community, including LINKAGE (Terwilliger and Ott 1994), SAGE (SAGE 1998), MAPMAKER/SIBS (Kruglyak and Lander 1995), GENEHUNTER (Kruglyak et al. 1996), and SIMWALK2 (Sobel and Lange 1996). For example, in the presence of incomplete information, Kruglyak et al. (1996) implemented a "perfect-data approximation method" to estimate significance levels in GENEHUNTER. However, Kong and Cox (1997), noting that this approximation might be unacceptably conservative, proposed and implemented a one-parameter model that allows exact calculations of likelihood ratios in GENEHUNTER-PLUS. They also noted that, when the information is far from complete, obtaining a good approximation without extensive simulation is difficult.

Simulation methods are often used in human linkage analyses as a substitute for analytical calculations that are too complex to be done (Ott 1991). They have been proposed as a means to study Hardy-Weinberg equilibrium and linkage equilibrium (Guo and Thompson 1992; Long et al. 1995; Slatkin and Excoffier 1996; Lazzeroni and Lange 1997; Zhao et al. 1999), to test disease-marker associations (Sham and Curtis 1995), to examine transmission disequilibrium (Morris et al. 1997; Lazzeroni and Lange 1998), to predict the maximum LOD score from pedigrees with known phenotypes (Boehnke 1986; Ploughman and Boehnke 1989), to estimate genetic risks (Sandkuyl and Ott 1989), and to approximate the statistical significance level for an observed test statistic (Ott 1989; Weeks et al. 1990; Davis et al. 1996; Sobel and Lange 1996; Sawcer et al. 1997; Kruglyak and Daly 1998; Guerra et al. 1999). Daniels et al. (1996) recently used the simulation approach in a genomewide search of quantitative-trait loci (QTL) underlying asthma. Generally, genotypes of each individual in the pedigrees are simulated, and each simulated sample is subject to the same analysis to derive an empirical distribution of the test statistic. Although simulation of genotypes is relatively straightforward, it can be very time-consuming when simulations are done conditional on partial pedigree information (Davis et al. 1996). In addition, if linkage analysis is performed with use of all markers on the same chromosome, such as in the method implemented in GENEHUNTER, calcula-

tion of inheritance vector probabilities can also be time-consuming.

Rather than the simulation of genotypes, permutation methods have been proposed as a means to derive an empirical significance level to map QTL in model organisms (Churchill and Doerge 1994; Doerge and Churchill 1996), as well as in human sib pairs (Wan et al. 1997). In these permutation tests, either the trait values or the trait differences between sibs are permuted, whereas the observed genotypes or the observed allele-sharing between sib pairs is fixed for each permutation. If certain environmental factors are known to have major effects on the trait, all of the trait values have to be adjusted by these factors before the residuals are permuted. However, for pedigrees having more-complex structures, such as those both with sibships and with other relative sets, it is not apparent what will be permuted, because of the lack of a uniform structure across the pedigrees. Furthermore, the permutation method discussed by Churchill and Doerge (1994) and by Wan et al. (1997) is not applicable to certain study designs and/or test statistics (such as pedigrees consisting of affected sib pairs and their parents).

Because of the limitations of the existing simulation methods—both those that are done on the basis of simulation of genotypes and those that are done on the basis of simulation of trait values—we propose a novel simulation method to estimate the significance level of the observed test statistic. The basic idea of this new method is that, if there is no linkage between a trait and a locus, then both grandpaternal and grandmaternal marker alleles in one parent are equally likely to be transmitted to the offspring. In our randomization procedure, we assign grandpaternal and grandmaternal alleles in one parent to his or her offspring, in such a way that each simulated sample is equally likely to occur under the null hypothesis of no linkage. In contrast to the methods that are done on the basis of simulation of genotypes, this approach does not need to regenerate each individual's genotypes, thus reducing the computation time for linkage analysis with use of all markers on the same chromosome. Compared with permutation tests, our new approach is applicable to arbitrary pedigree structures and to both qualitative and quantitative traits, provided that inheritance vector probabilities can be calculated or estimated. Results from our simulation studies suggest that the randomization method generally has the correct rate of type 1 error in linkage analysis. This randomization method has been implemented in a modified version of GENEHUNTER. For pedigrees of moderate sizes, GENEHUNTER can calculate the exact inheritance vector probabilities. However, the space required for all inheritance vectors is large in large pedigrees. It may be prohibitive to calculate all inheritance vector probabilities, and simulation-based methods may be used to estimate the inheritance vector distribution. Our novel simulation method is described in the Methods section and is evaluated in the Simulations and Application sections.

## Methods

In this section, we describe how the randomization procedure generates simulated samples that are then used to derive an empirical distribution of any test statistic. Let $\mathbf{Y} = (y_1, y_2, \ldots, y_{n+f})$ denote the trait values of $n + f$ pedigree members, in a pedigree with $f$ founders (i.e., those individuals whose parents are not in the pedigree) and $n$ nonfounders (i.e., those individuals whose parents are in the pedigree). The trait can be either qualitative or quantitative, and the trait values may be missing for some individuals. Suppose that a subset of the members in the pedigree has been typed for a set of markers. The inheritance pattern at each locus $x$ is completely described by a binary inheritance vector $\mathbf{v}(\mathbf{x}) = (f_1, m_1, f_2, m_2, \ldots, f_n, m_n)$, where $f_i = 0$ or 1 if the grandpaternal or grandmaternal allele was transmitted to the $i$th nonfounder from its father, and $m_i = 0$ or 1 if the grandpaternal or grandmaternal allele was transmitted to the $i$th nonfounder from its mother (Lander and Green 1987). There are a total of $2^{2n}$ possible inheritance vectors for a pedigree with $n$ nonfounders. In general, the actual inheritance vector cannot be uniquely determined from the marker data, and we need to estimate the probability of each inheritance vector. There are efficient algorithms to calculate the inheritance vector probabilities for pedigrees of moderate size (Whittemore and Halpern 1994*b;* Kruglyak et al. 1996). The inheritance vector probabilities are independent of the trait values of the people in the pedigree.

The rationale of our proposed randomization procedure is that, if a locus is not linked to the trait locus, then the grandpaternal and grandmaternal alleles should have an equal chance of transmittal to the offspring. Therefore, $f_i$ and $m_i$, in the inheritance vector, have an equal chance to be 0 or 1 in the $i$th nonfounder in the pedigree. We describe our simulation procedure, in order, for single individuals, two full sibs, and general pedigrees.

### *One Marker, Single Individuals*

For a pedigree with only one nonfounder, the inheritance vector has two components ($f$, $m$), with four possibilities: (0,0), (0,1), (1,0), and (1,1). If the inheritance vector can be uniquely determined, then the randomization procedure proceeds as follows. We first generate an indicator vector $(r_f, r_m)$, where $r_f$ and $r_m$ have an equal chance to be 0 or 1. The $r_f$ is the indicator of whether, in the simulated sample, the same grandparental allele

would be transmitted to this individual, from his or her father, as is transmitted in the observed sample. If $r_f$ is 0, then the same grandparental allele will be present in the offspring, and if $r_f$ is 1, then the other grandparental allele will be present in the offspring. The $r_m$ is similarly defined for the transmission from the mother to this person. It is easy to see that the randomization procedure has an equal chance to generate four inheritance vectors: (0,0), (0,1), (1,0), and (1,1).

When there are uncertainties in inferring the exact inheritance vector, let $P_{(00)}$, $P_{(01)}$, $P_{(10)}$, and $P_{(11)}$ denote the probabilities for inheritance vectors (0,0), (0,1), (1,0), and (1,1), respectively. From each indicator vector ($r_f$, $r_m$), each randomization produces a set of new inheritance vector probabilities $P^r_{(fm)}$, as follows. If the indicator vector ($r_f$, $r_m$) is (0,0), then the new set of inheritance vector probabilities is the same as those observed. If the indicator vector is (0,1), then the new set of inheritance vector probabilities is created by the retention of the $f$ values in the inheritance vectors but change of the $m$ values from 1 to 0 or from 0 to 1. This results in $P^r_{(00)} = P_{(01)}$, $P^r_{(01)} = P_{(00)}$, $P^r_{(10)} = P_{(11)}$, and $P^r_{(11)} = P_{(10)}$, where the $P^r_{(fm)}$ are the new inheritance vector probabilities. A new set of inheritance vector probabilities can be similarly defined if the indicator vector ($r_f$, $r_m$) is (1,0) or (1,1). If we use the modular arithmetic notation, we can generally write $P^r_{(fm)} = P_{(f^r,m^r)}$, where $f^r = (f + r_f) \pmod 2$, and $m^r = (m + r_m) \pmod 2$. The notation "$a \pmod N$" represents the remainder of $a$ divided by $N$. For addition modulo 2, $0 \pmod 2 = 0$, $1 \pmod 2 = 1$, and $2 \pmod 2 = 0$.

### *One Marker, Two Full Sibs*

There are a total of 16 possible inheritance vectors for two full sibs, $(f_1,m_1,f_2,m_2)$, and the allele-sharing probabilities between two sibs can be derived from these probabilities. For example, the probability that the sib pair share 0 alleles identical by descent (IBD) is $P_{(1100)} + P_{(1001)} + P_{(0110)} + P_{(0011)}$.

Suppose that the inheritance vector can be uniquely determined, and, without loss of generality, suppose that $(f_1,m_1,f_2,m_2) = (0000)$. We first generate an indicator vector $\mathbf{R} = (r_{f_1},r_{m_1},r_{f_2},r_{m_2})$, with each component having an equal chance to be 0 or 1. The $r_{f_i}$ is the indicator of whether the same grandparental allele would be transmitted to the $i$th sib from its father in the simulated sample as was transmitted in the observed sample. The $r_{m_i}$ is similarly defined as the indicator for the transmission from the mother to the $i$th sib. With the 16 possibilities for $\mathbf{R}$, the randomization procedure has an equal chance to generate all 16 inheritance vectors for $(f_1,m_1,f_2,m_2)$. Thus, the probabilities that the sib pair share 0, 1, and 2 alleles IBD in a simulated sample are $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$, respectively.

If the true inheritance vector cannot be uniquely determined, then the probability for inheritance vector $(f_1,m_1,f_2,m_2)$ can be represented by $P_{(f_1,m_1,f_2,m_2)}$. The randomization procedure simulates samples as follows: (1) Randomly generate an indicator vector $(r_{f_1},r_{m_1},r_{f_2},r_{m_2})$, with each component having an equal chance to be 0 or 1. (2) Define a set of "randomized" inheritance vector probabilities $P^r_{(f_1,m_1,f_2,m_2)} = P_{(f^r_1,m^r_1,f^r_2,m^r_2)}$, where $f^r_1 = f_1$ if $r_{f_1} = 0$, $f^r_1 = 1 - f_1$ if $r_{f_1} = 1$, and $m^r_1$,$f^r_2$, and $m^r_2$ are similarly defined (i.e., $f^r_i = f_i + r_{f_i} \pmod 2$ and $m^r_i = m_i + r_{m_i} \pmod 2$).

### *One Marker, General Pedigrees*

The generalization of this randomization procedure to pedigrees of arbitrary structure is straightforward. For an arbitrary pedigree with $n$ nonfounders, denote the probability of the inheritance vector $(f_1,m_1,f_2,m_2,\ldots, f_n,m_n)$ by $P_{(f_1,m_1,f_2,m_2,\ldots,f_n,m_n)}$. Each randomization uses an indicator vector $\mathbf{R} = (r_{f_1},r_{m_1},r_{f_2},r_{m_2},\ldots,r_{f_n},r_{m_n})$ to generate a new set of inheritance vector probabilities $P^r_{(f_1,m_1,f_2,m_2,\ldots,f_n,m_n)} = P_{(f^r_1,m^r_1,f^r_2,m^r_2,\ldots,f^r_n,m^r_n)}$, where $f^r_i = f_i$ if $r_{f_i} = 0$, $f^r_i = 1 - f_i$ if $r_{f_i} = 1$, and $m^r_i$ is similarly defined. Statistical tests can be performed on simulated samples to derive an empirical distribution under the assumption of no linkage.

For a locus $s$ on a given chromosome, several measures have been proposed, in the literature, to summarize the uncertainty in inheritance vectors at this locus (Kruglyak and Lander 1995; Kruglyak et al. 1996; Teng and Siegmund 1998). For the entropy measure $I_E(s) = -\Sigma P_i \log_2 P_i$, introduced by Kruglyak et al. (1996), where the $P_i$ are the inheritance vector probabilities, each randomized sample preserves the amount of genetic information measured by $I_E(s)$ at each locus, i.e., $I_E(s)$ in each simulated sample is the same as that in the observed sample. This is so because, in each simulated sample, the simulated inheritance vector probabilities at each locus are some permutation of the observed inheritance vector probabilities.

### *Multiple Markers, General Pedigrees*

In the above discussion, only a single marker was considered in the randomization procedure. A genomewide scan usually involves several hundred markers in the genome. If we are interested only in determining pointwise statistical significance levels, we may simply apply the randomization procedure to each point along the chromosome to estimate statistical significance. However, hundreds of markers are generally screened in a genomewide scan, and there is a need to control the genomewide false-positive rate when these markers, together with chromosomal locations between these markers, are included in a single linkage study.

Because markers on the same chromosomes are dependent, both theoretical (Feingold et al. 1993; Teng and Siegmund 1998) and simulation (Churchill and Doerge 1994) methods have been proposed to take this dependence into account to determine genomewide threshold values. With the simulation procedure, we can use the test statistics calculated from all of the markers, in the simulated samples, in different ways to assist our inference of statistical significance. For example, to determine the genomewide statistical significance for the largest observed test statistic $T_{max}$, we may keep track of the largest test statistic, $T^i_{max}$, from the $i$th simulated sample and may estimate the genomewide significance level from the proportion of times that simulated $T^i_{max}$ is larger than the observed $T_{max}$. Note that, if we focus on the locus that attains the largest statistic $T_{max}$, we may miss "secondary" loci. An alternative approach is to decide on a threshold $t$ for the genomewide statistical significance level $\alpha$ and then declare that all loci yielding observed test statistics larger than $t$ are statistically significant. Under the simulation procedure, we can set this genomewide threshold as the $100(1 - \alpha)$ percentile for all of the simulated test statistics.

If, in each simulation, we use different indicator vectors to generate simulated samples for markers on the same chromosome, then the dependent structure of these markers will be lost. To maintain the dependence among markers on the same chromosome, for each simulated sample, we need to use the same indicator vector to generate new inheritance probabilities for all of the markers. By application of the same indicator vector to all markers on the same chromosome, the amount of recombination and the information on where crossovers have/might have occurred in each simulated sample remain the same as in the observed sample, thereby maintaining the dependent structure among these markers.

## Implementation

The proposed simulation procedure has been implemented in a modified version of GENEHUNTER, to estimate the statistical significance for linkage at each position along a specific chromosome. Because both GENEHUNTER and GENEHUNTER-PLUS produce a $Z$ score to summarize the statistical significance at each individual point, our modified program also calculates a similar $Z$ score, as follows. First, a normalized test statistic $Z_i = (T - \mu)/\sigma$ is calculated for each pedigree, where $\mu$ and $\sigma$ are the mean and SD of the simulated test statistics, respectively. Then, the overall $Z$ score is calculated as a weighted sum of these normalized test statistics (i.e., $Z = \Sigma w_i Z_i$, where $\Sigma w_i^2 = 1$). In the following analyses, all the weights are set to be equal (i.e., $w_i = 1/\sqrt{N}$), where $N$ is the number of pedigrees in the sample. When the number of families is large, the observed $Z$ score can be compared with a standard normal distribution, to estimate the statistical significance level. Although our simulation studies reveal that this standardization is best suited to nuclear families, there may be potential bias for complex pedigree structures, such as pedigrees involving three or more generations. Another source of possible bias is missing genotypes that cannot be inferred from other people in the pedigree. A more accurate way to estimate overall significance levels, from all pedigrees, is through the convolution of the empirical distributions of all the families, especially when the sample size is small. We will make the modified GENEHUNTER program available for the genetics community after the interface is improved.

## Simulations

In this section, we apply the simulation procedure to the simulated data on a complex trait from Genetic Analysis Workshop 10 (GAW10) problem 2A (MacCluer et al. 1997). GAW10 problem 2 involves a simulated common disease defined by imposing a threshold, T, on a quantitative trait, Q1. Every individual with a value of $Q1 \geqslant 40$ is defined as "affected." Q1 is associated with four other quantitative traits (Q2–Q5) and an environmental factor. There are six major genes influencing one or more of the five quantitative traits (Q1–Q5). There is one major gene (*MG1*) on chromosome 5 that accounts for 21% of the variance for Q1. MacCluer et al. (1997) gave a complete description of the generating model. Problem 2A consisted of 200 replicates of 239 nuclear families containing 1,164 individuals. These pedigrees were randomly ascertained, subject to the constraint that there be at least two living offspring. A total of 367 highly polymorphic markers, spaced an average of 2.03 cM apart on 10 chromosomes, were available for each individual.

Among the first 100 replicates, there were 576 families with at least two affected individuals each. We used GENEHUNTER-PLUS and the modified program to analyze these 576 families. There are two scoring functions of nonparametric linkage in GENEHUNTER, $NPL_{pairs}$ and $NPL_{all}$. $NPL_{pairs}$ simply calculates the number of pairs of alleles from distinct affected pedigree members that are IBD. $NPL_{all}$, introduced by Whittemore and Halpern (1994*a*), puts extra weight on three or more affected pedigree members who are IBD. We used $NPL_{all}$ for our analysis. For each of the methods, all of the families are assigned the same weight, to obtain the overall test statistic $Z$ at the markers. We compare the results from three different methods: GENEHUNTER, GENEHUNTER-PLUS, and the modified GENEHUNTER program using the randomization test. Each of these three programs summarizes the statistical significance against the null hypothesis of no linkage with an overall

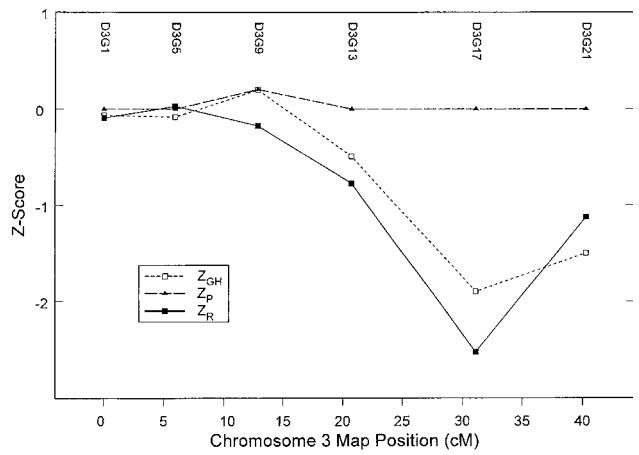


**Figure 1** Overall $Z$ scores calculated with GENEHUNTER ($Z_{GH}$), GENEHUNTER-PLUS ($Z_P$), and the randomization procedure ($Z_R$), at six markers on chromosome 3, for the simulated data from GAW10 problem 2A. A total of 576 pedigrees with two or more affected individuals each were analyzed.

$Z$ score. The significance level can be estimated with use of $1 - \Phi(Z)$, where $\Phi(Z)$ is the cumulative distribution for the standard normal distribution. We use "$Z_{GH}$," "$Z_P$," and "$Z_R$" to denote the overall test statistics calculated, respectively, with use of GENEHUNTER, GENEHUNTER-PLUS, and the randomization procedure. For the randomization test, 500 randomized samples are simulated. When all of the markers are used, the genetic information is almost complete throughout the genome, and all three methods give very similar results (data not shown).

In practice, the first round of a genomewide scan usually involves less densely distributed markers. In figures 1 and 2, we show the results from an analysis using six markers from chromosome 3 (with average distance of 8.06 cM) and seven markers from chromosome 5 (with average distance of 7.1 cM). There is no gene on chromosome 3 associated with the disease, and all three methods give negative results (fig. 1). There is one major gene (*MG1*), located 28.7 cM from the left end on chromosome 5. All three methods yield strong signals for a gene located in the interval *D5G13–D5G17–D5G21* (fig. 2). The maximum $Z$ scores are $Z_R = 6.33$, $Z_P = 6.10$, and $Z_{GH} = 5.80$, with corresponding pointwise $P$ values of $1.2 \times 10^{-10}$, $5.3 \times 10^{-10}$, and $3.3 \times 10^{-9}$.

## Application

Type 1 diabetes, or insulin-dependent diabetes mellitus (IDDM), is a complex disorder, in which both genetic and environmental factors contribute to the development of the disease. A genomewide scan in affected sib pairs identified *IDDM1* (in the major histocompatibility complex on chromosome 6p21), *IDDM2* (in the insulin-gene region on chromosome 11p15), and 10 other chromosomal regions with some positive evidence (i.e., $P <$ .005) of linkage to IDDM (Davies et al. 1994). Two of these 10 regions are on chromosome 6q, near the markers *ESR* and *D6S264,* and they are named "*IDDM5*" and "*IDDM8,*" respectively. Subsequent studies confirmed these two susceptibility genes for IDDM (Davies et al. 1996; Luo et al. 1996; Delepine et al. 1997; Mein et al. 1998).

In this section, we apply our modified GENEHUNTER program to analyze IDDM data on chromosome 6q, reported by Davies et al. (1996). These data were downloaded from the World Wide Web site maintained by the Todd group at Cambridge University (Index of /todd/HumanData/chr6). Among the 299 U.K. pedigrees examined by Davies et al., 285 are available at their Web site. Each pedigree has both parents and two affected sibs. A total of 39 markers were studied in these pedigrees, to confirm the positive findings from their previous study (Davies et al. 1994). To mimic a typical genomewide scan scenario, we consider only the 11 markers studied in the initial genome scan done by Davies et al. (1994)—namely, *D6S308, D6S314, D6S310, D6S311, ESR, D6S290, D6S441, D6S415, D6S305, D6S264,* and *D6S281.*

For the affected-sib-pair families, the two scoring methods in GENEHUNTER—$NPL_{pairs}$ and $NPL_{all}$—are equivalent. For each of the methods, all of the families were assigned the same weight, to obtain the overall test statistic $Z$ at the 11 markers. For the randomization test, 500 randomized samples were simulated. In figure 3, we

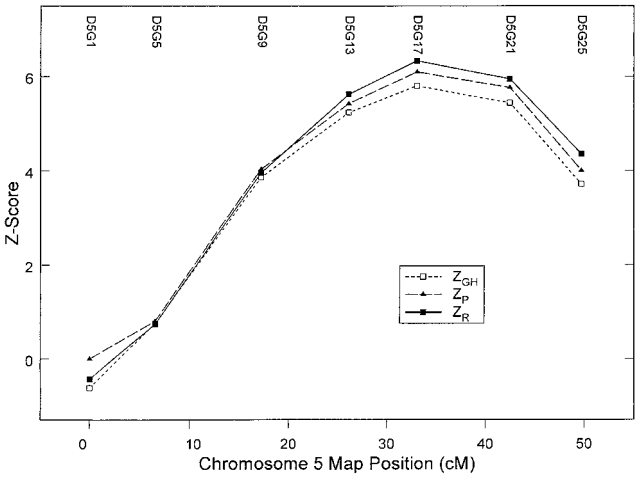


**Figure 2** Overall $Z$ scores calculated with GENEHUNTER ($Z_{GH}$), GENEHUNTER-PLUS ($Z_P$), and the randomization procedure ($Z_R$), at seven markers on chromosome 5, for the simulated data from GAW10 problem 2A. A total of 576 pedigrees with two or more affected individuals each were analyzed.

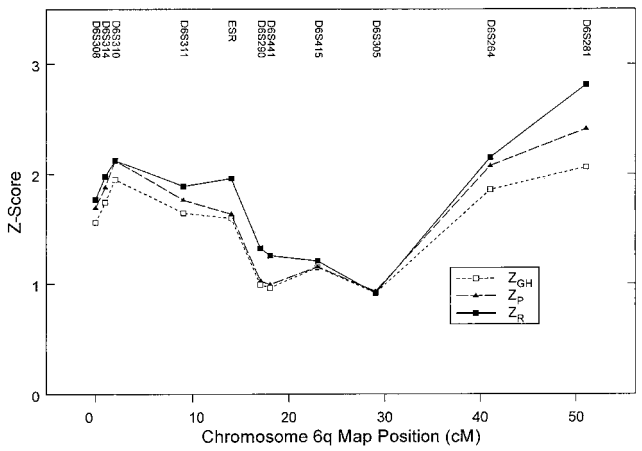


**Figure 3** Observed overall $Z$ scores calculated with GENEHUNTER ($Z_{GH}$), GENEHUNTER-PLUS ($Z_P$), and the randomization procedure ($Z_R$), at 11 markers on the long arm of chromosome 6. A total of 285 U.K. diabetes pedigrees were analyzed.

plot the $Z$ scores from these three methods. For this particular data set, $Z_R > Z_P > Z_{GH}$ across all of the markers; therefore, the statistical significance levels for linkage that are obtained with the randomization procedure are always smaller than those that are obtained with either GENEHUNTER or GENEHUNTER-PLUS. The largest $Z$ score is obtained at *D6S281,* for all three methods, as follows: $Z_R = 2.81$, $Z_P = 2.41$, and $Z_{GH} = 2.06$, respectively; the corresponding pointwise $P$ values are .002, .008, and .02.

Note that the strict inequality $Z_R > Z_P > Z_{GH}$ among these three methods does not hold for all data sets; for example, on the basis of its own definition, $Z_P$ cannot be negative, whereas $Z_R$ can be <0 when the allele sharing among the affected relatives is less than what is expected under the null hypothesis of no linkage. However, both for the IDDM data that we analyze here and for an alcoholism data set that we have analyzed for the Genetic Analysis Workshop 11 (Zhao et al., in press), when $Z_R > 0$, $Z_R > Z_P$, in most cases, leading to more-significant findings at each individual marker locus, with use of the randomization test.

This diabetes data set was analyzed on a DIGITAL Alpha 5/300 workstation. The running time for GENEHUNTER-PLUS was 221 s, and the running time for the new simulation procedure was 231 s. The extra time needed to perform the randomization test was minimal when compared with that needed to calculate the inheritance vector probabilities for all of the pedigrees in the sample.

### Discussion

Compared with the existing simulation methods, our proposed simulation procedure has several advantages: (1) it is applicable to arbitrary pedigree structures, provided that the inheritance vector probabilities can be calculated or estimated; (2) it can be applied to study both qualitative and quantitative traits; and (3) computation is minimal, once the inheritance vector probabilities have been estimated in the original sample. In addition, the rate of type I error has been found to be near the nominal level in our simulation studies.

With use of the same test statistic, the randomization procedure and other methods differ only in the estimation of the statistical significance. Our simulation studies showed that this randomization procedure has better power than some methods (e.g., the perfect-data approximation method in GENEHUNTER) that suffer from reduced power because of the conservative nature of their testing procedures.

In our studies using simulated data, the randomization procedure did as well as or better than alternative methods in the assessment of the statistical significance. Our proposed randomization test is based on a set of inheritance vector probabilities. Because our current implementation is limited to GENEHUNTER, we used only pedigrees of moderate size to map genes for qualitative traits. For large pedigrees, the space of all inheritance vectors is large, and it may be prohibitive to calculate all inheritance vector probabilities. Simulation-based programs, such as SIMWALK2 (Sobel and Lange 1996), may be used to estimate the inheritance vector distribution. Similarly, our simulation procedure can be implemented to map QTL, with use of pedigrees of arbitrary structures.

To facilitate comparison with GENEHUNTER and GENEHUNTER-PLUS, we calculated a normalized test statistic $Z$ for each pedigree by applying our simulation method to nuclear families. However, application of this standardization to each pedigree may result in bias for pedigrees with three or more generations. A more direct estimate of the significance level is obtained by the comparison of the observed test statistic summed over all pedigrees, with the convolution of the empirical distributions for the test statistic from these pedigrees. In addition, many missing genotypes in a pedigree may also lead to possible bias in the estimation of the statistical significance. Because of these possible biases in our randomization procedure, our method must be considered an approximate simulation method. Although the conditions under which the proposed simulation procedure is valid are currently being explored, simulation studies have demonstrated that this procedure generally has approximately the correct nominal false-positive rates and should prove useful for linkage studies.

## Acknowledgments

We thank Drs. Charles C. Berry, Robert C. Elston, Augustine C. Kong, Andrew J. Pakstis, and Fred Wright and two anon-

ymous referees for their constructive comments. This work was supported in part by grants GM59507, HD36834, MH30929, MH39239, GM57672, and DA09055 from the National Institutes of Health (NIH). We thank Dr. MacCluer for providing us the simulated data from GAW10. GAW is supported by NIH grant GM31575.

## Electronic-Database Information

The URL for data in this article is as follows:

Index of /todd/HumanData/chr6, http://diesel.cimr.cam.uk.ac/todd/HumanData/chr6 (for IDDM data on chromosome 6q)

## References

Boehnke M (1986) Estimating the power of a proposed linkage study: a practical computer simulation approach. Am J Hum Genet 39:513–527

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

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

Daniels SE, Bhattacharrya S, James A, Leaves NI, Young A, Hill MR, Faux JA, et al (1996) A genome-wide search for quantitative trait loci underlying asthma. Nature 383: 247–250

Davies JL, Cucca F, Goy JV, Atta ZA, Merriman ME, Wilson A, Barnett AH, et al (1996) Saturation multipoint linkage mapping of chromosome 6q in type I diabetes. Hum Mol Genet 5:1071–1074

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, et al (1994) A genome-wide search for human type I diabetes susceptibility genes. Nature 371: 130–136

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

Delepine M, Pociot F, Habita C, Hashimoto L, Froguel P, Rotter J, Cambon-Thomsen A, et al (1997) Evidence of a non-MHC susceptibility locus in type I diabetes linked to HLA on chromosome 6. Am J Hum Genet 60:174–187

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

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

Guo SW, Thompson E (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics 48:361–372

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

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

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

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

Lander ES, Green P (1987) Construction of multilocus genetic maps in humans. Proc Natl Acad Sci USA 84:2363–2367

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

Lazzeroni LC, Lange K (1997) Markov chains for Monte Carlo tests of genetic equilibrium in multidimensional contingency tables. Ann Stat 25:138–168

——— (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48: 67–81

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799–810

Luo D-F, Buzetti R, Rotter JI, Maclaren N, Raffel L, Nistico L, Giovannini C, et al (1996) Confirmation of three susceptibility genes to insulin-dependent diabetes mellitus: *IDDM4, IDDM5,* and *IDDM8*. Hum Mol Genet 5: 693–698

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

Mein CA, Esposito L, Dunn MG, Johnson GCL, Timms AE, Goy JV, Smith AN, et al (1998) A search for type 1 diabetes susceptibility genes in families from the United Kingdom. Nat Genet 19:297–300

Morris AP, Curnow RN, Whittaker JC (1997) Randomization tests of disease-marker associations. Ann Hum Genet 61: 49–60

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

——— (1991) Analysis of human genetic linkage, rev ed. Johns Hopkins University Press, Baltimore

Ploughman LM, Boehnke M (1989) Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet 44:543–551

SAGE (1998) Statistical analysis for genetic epidemiology. Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland

Sandkuyl LA, Ott J (1989) Determining informativity of marker typing for genetic counseling in a pedigree. Hum Genet 82:159–162

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

Sham PC, Curtis D (1995) Monte Carlo tests for associations between disease and alleles at highly polymorphic loci. Ann Hum Genet 59:97–105

Slatkin M, Excoffier L (1996) Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm. Heredity 76:377–383

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323–1337

Teng J, Siegmund D (1998) Multipoint linkage analysis using affected relative pairs and partially informative markers. Biometrics 54:1247–1265

Terwilliger JD, Ott J (1994) Handbook of human linkage analysis. Johns Hopkins University Press, Baltimore

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

Weeks DE, Ott J, Lathrop GM (1990) SLINK: a general simulation program for linkage analysis. Am J Hum Genet Suppl 47:A204

Whittemore AS, Halpern J (1994*a*) A class of tests of linkage using affected pedigree members. Biometrics 50:118–127

——— (1994*b*) Probability of gene identity by descent: computation and applications. Biometrics 50:109–117

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

Zhao H, Pakstis AJ, Kidd JR, Kidd KK (1999) Assessing linkage disequilibrium in a complex genetic system. I. Overall deviation from random association. Ann Hum Genet 63: 167–179

Zhao H, Sheffield LJ, Pakstis AJ, Knauert MP, Kidd KK. A more powerful method to evaluate p-values in GENEHUNTER. In: Goldin L, Amos CI, Chase GA, Goldstein AM, Jarvik GP, Martinez MM, Suarez BK, et al (eds) Genetic Analysis Workshop 11: analysis of genetic and environmental factors in common diseases. Genet Epidemiol (in press)