RESEARCH ARTICLE

An estimating function approach to linkage heterogeneity

HE GAO[1,2], YING ZHOU[1*], WEIJUN MA[1], HAIDONG LIU[1] and LINAN ZHAO[1]

[1]*Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, People's Republic of China*
[2]*School of Mathematical Sciences, Harbin Normal University, Harbin 150025, People's Republic of China*

Abstract

Testing linkage heterogeneity between two loci is an important issue in genetics. Currently, there are four methods (K-test, A-test, B-test and D-test) for testing linkage heterogeneity in linkage analysis, which are based on the likelihood-ratio test. Among them, the commonly used methods are the K-test and A-test. In this paper, we present a novel test method which is different from the above four tests, called G-test. The new test statistic is based on estimating function, possessing a theoretic asymptotic distribution, and therefore demonstrates its own advantages. The proposed test is applied to analyse a real pedigree dataset. Our simulation results also indicate that the G-test performs well in terms of power of testing linkage heterogeneity and outperforms the current methods to some degree.

[Gao H., Zhou Y., Ma W., Liu H. and Zhao L. 2013 An estimating function approach to linkage heterogeneity. *J. Genet.* **92**, 413–421]

Introduction

Genetic heterogeneity is divided into allelic heterogeneity and nonallelic heterogeneity. Nonallelic heterogeneity is also known as locus heterogeneity or linkage heterogeneity, and refers to the case where the same disease trait can be caused by the genes on different loci (Burmeiste *et al.* 2008). From the perspective of linkage, it is the inconsistency or inequality of linkage/recombination degree between a marker and a trait locus to be tested on the same chromosome among independent individuals or pedigrees. It is well known that recombination rate can be seen as a measure of distance between two linked loci and gene mapping can be carried out by examining the linkage degree between a marker and a trait locus in linkage analysis. However, the presence of linkage heterogeneity would obviously impact the results and efficiency of linkage analysis, and therefore would decrease the accuracy of gene mapping. Of course, the effective inference on linkage heterogeneity can help people to better understand complex genetic traits such as many diseases. Therefore, testing whether there exists linkage heterogeneity between two loci of interest is very important in biology, genetics and medical sciences (Lewis *et al.* 1993; Huber *et al.* 2009; Bisceglia *et al.* 2010; McClellan and King 2010).

Aiming at linkage heterogeneity between the two loci, researchers have proposed several test methods, in which there are four basic methods, i.e., the K-test (Morton 1956), the A-test (Smith 1963; Ott 1977; Risch and Baron 1982), the B-test (Risch 1988) and the D-test (Darlene 1994). The four methods are based on the idea of the likelihood-ratio test. The forms of these test statistics are, respectively,

$$Z_K = 2\left[\sum_{i=1}^{F}\max_{\theta i}\ln L_i(\theta_i) - \max_{\theta_H}\sum_{i=1}^{F}\ln L_i(\theta_H)\right],$$

$$Z_A = 2\left[\max_{\theta_A,\alpha}\sum_{i=1}^{F}\ln L_i(\theta_A,\alpha) - \max_{\theta_H}\sum_{i=1}^{F}\ln L_i(\theta_H,\alpha=1)\right],$$

$$Z_B = 2\left[\max_{u_B,d_B}\sum_{i=1}^{F}\ln L_i(u_B,d_B) - \max_{u_H}\sum_{i=1}^{F}\ln L_i(u_H,d_B=0)\right],$$

$$Z_D = 2\left[\max_{u_D,d_D,\alpha}\sum_{i=1}^{F}\ln L_i(u_D,d_D,\alpha) - \max_{u_H}\sum_{i=1}^{F}\ln L_i(u_H,d_D=0,\alpha=1)\right].$$

The K-test allows each pedigree to have a different recombination rate θ_i ($i = 1, 2, \ldots, F$), where θ_i can be different from each other. There is no special requirement for the case of linkage heterogeneity in this method. The requirement of the A-test is special. It assumes that the pedigrees to be tested are divided into two classes: one class of pedigree is homogeneously linked between the considered loci,

*For correspondence. E-mail: yzhou@aliyun.com.

Keywords. linkage heterogeneity; likelihood ratio; power; recombination fraction; test.

sharing the same recombination rate; the other class of pedigree is not linked between the loci. The ratio of the two classes of pedigrees is α: (1–α), where $\alpha = 1$ means that the linkage situation of the tested pedigree is homogeneous. The A-test is also known as mixture test (Hodge *et al.* 1983). The B-test involves the idea of Bayesian theory. For all pedigrees, it assumes that there is a prior distribution $f(\theta; u_B, d_B)$ on recombination rate θ, where u_B ($0 < u_B < 0.5$) is the expectation, v_B is the variance and $d_B = v_B/\max v(u_B)$. In the D-test, it is assumed that the two considered loci are linked in some pedigree, with the recombination rate θ having a prior distribution $f_D(\theta; u_D, d_D)$; but in the rest pedigrees, the two loci are not linked and the proportion of the two pedigrees is α: (1−α). Obviously, the A-test and B-test are special cases of the D-test, since the D-test is just a combination of the A-test and B-test. For the B-test and D-test, $d_B = 0$ or $d_D = 0$ is equivalent to $P(\theta = u_H) = 1$, which means that the linkage situation is homogeneous. So the values of the four test statistics in the homogeneous parts (the second part in each statistic) are equivalent for the same data.

In these four tests, estimating parameters is an essential part. In fact, the parameters to be estimated under the null hypothesis are the same one, i.e., the homogeneous recombination rate θ_H, whereas the parameters under the alternative hypothesis of heterogeneity situation are different in the four techniques. The recombination rate θ_i ($i = 1, 2, \ldots, F$) for each pedigree needs to be estimated in the K-test; the proportion α of linkage families and shared recombination rate θ_H need to be estimated in the A-test; the expectation u_B of the prior distribution and the ratio d_B need to be estimated in the B-test; and proportion α, priori mean u_D and ratio d_D need to be estimated in the D-test. Usually the calculations involved in estimating parameters are tedious in the B-test and D-test.

Another problem is the distribution of the statistic in the test. Under the null hypothesis of homogeneity, the asymptotic distributions of the four test statistics are different. The asymptotic distribution of the K-statistic is a χ^2 distribution with F-1 degrees of freedom (Morton 1956); the asymptotic distributions of the A-statistic and B-statistic are the mixture between χ_0^2 (degraded single-point distribution) and χ_1^2 with the ratio of 50:50 (Ott 1999); and the asymptotic distribution of the D-statistic is the mixture distribution of χ_0^2, χ_1^2 and χ_2^2, with the proportion of $0.5 - P_D : 0.5 : P_D$ ($0 \le P_D \le 1/2$). In fact, the asymptotic distribution of the D-test is difficult to obtain (Darlene 1994).

Risch (1988) pointed out that the statistical distribution of the K-test is sensitive to the number of pedigrees, while the A-test and B-test are not sensitive to the number. All three statistics are slightly sensitive to θ. Moreover, simulation results show that the B-test is a little better, but there are slight differences among the three tests in fact. Darlene (1994) further pointed out that there is no best method on testing linkage heterogeneity in any situation. Based on the simulated results of the above two researchers, for tight linkage, the K-test is conservative, but it is often not conservative for medium and loose linkage. The A-test and B-test are conservative in most cases and the smaller the pedigree structure, higher the level of conservativeness will be. The D-test is completely conservative and the hypothesis of heterogeneity will not be easily accepted. At the same time, the powers are sensitive to the number of pedigrees and the degree of linkage heterogeneity. Leal and Ott (2000) discussed the problem of trade-off between power and sample size in genetic heterogeneity studies.

In practice, many questions of linkage heterogeneity about disease loci always accompany the methods of the K-test and A-test (Stine *et al.* 1995; Buxbaum *et al.* 2001; Grice *et al.* 2002; Hauser *et al.* 2004). The applications of B-test and D-test are relatively rare because of the tedious parameter estimation and their distribution problems.

In this paper, by means of an estimating function, we present a new test method for the linkage heterogeneity problem, here referred to as G-test, the idea of which is different from that of the likelihood-ratio test. The G-test allows each pedigree to have its own recombination rate and contains the special situation of only involving two pedigrees such as in A-test. Comparing with the previous four methods, the G-test has its advantages in parameter estimation, the properties of distribution of the G-statistic and test power, etc. The new method can be used to analyse data from both experimental and human populations.

Background and method

First, we consider the problem of linkage heterogeneity for a backcross (BC) population under random mating. Let M and m denote alleles of a marker, and Q and q denote alleles of a trait locus. Suppose recessive inheritance at the trait locus, which means that the individuals with genotype qq are abnormal, while the individuals with QQ or Qq are normal. The backcross offspring have four possible phases: (i) MQ/mq, (ii) Mq/mq, (iii) mQ/mq and (iv) mq/mq.

Assuming that the frequencies of genotypes of markers can be measured, the number of the above four phases can be inferred because of recessive inheritance on the trait locus. Let the number of the considered pedigrees be F, the number of observed individuals in each family be n_i ($i = 1, 2, \ldots, F$), where $\sum_{i=1}^{F} n_i = n$ and the number of individuals with the jth phase in the ith pedigree be r_{ij} ($i = 1, 2, ..., F$; $j = 1, 2, 3, 4$). Let θ_i ($i = 1, 2, \ldots, F$) denote the recombination rate between the marker and the trait loci in the ith pedigree. Let estimating function

$$g_i(\theta_i) = \frac{d \ln L_i(\theta_i)}{d\theta_i} = \frac{r_{i2} + r_{i3}}{\theta_i} - \frac{r_{i1} + r_{i4}}{1 - \theta_i}, \quad (i = 1, 2, \ldots, F) \tag{1}$$

where $L_i(\theta_i)$ is the likelihood function for the ith pedigree. From equation (1), we can obtain

$$\mathrm{var}\,[g_i(\theta_i)] = \frac{n_i}{\theta_i(1 - \theta_i)}.$$

Under the condition of $H_0 : \theta_1 = \theta_2 = \cdots = \theta_F$, we let $\theta_i = \theta$ $(i = 1, 2, \ldots, F)$. Using the total observed data of the F pedigrees, we can obtain the MLE $\hat{\theta}$ of θ under H_0 as

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{F}(r_{i2} + r_{i3}).$$

Then, we construct the statistic as follows:

$$G = \sum_{i=1}^{F}\frac{\left[g_i(\hat{\theta})\right]^2}{\mathrm{var}\left[g_i(\hat{\theta})\right]}.$$

In the expression of the statistic G, $g_i(\theta_i)$ is the estimating (score) function of the recombination rate of the ith pedigree (it usually can be constructed by the observed likelihood function), $\mathrm{var}\left[g_i(\hat{\theta}_i)\right]$ is the variance of estimation function $g_i(\hat{\theta}_i)$, where $\hat{\theta}$ is the estimate of θ, which can be obtained from the observation data of all families.

The corresponding test is called the G-test. Note that the G-test only needs the estimate of θ under H_0, so the value of the statistic is easy to compute. Also, the statistic consists of the estimating function of parameter θ, which sufficiently includes the statistical information of θ. Further, it can be verified that the statistic G satisfies the conditions of the Lemma in the Appendix (Fujii 1994) due to concavity of log-likelihood and the famous central limit theorem, therefore, the asymptotic distribution of the statistic G is a linear combination of two independent chi-square distributions with the degrees of freedom $F-2$ and 1, respectively, i.e. $\chi^2(F-2) + c\chi^2(1)$, where c is a constant (see Appendix). Here, it can be verified that the coefficient c is equal to 1 for the recessive inheritance model and then the asymptotic distribution of the statistic G is $\chi^2(F-1)$. Based on the above reasons, we apply the G-statistic to test the problem of linkage heterogeneity.

In addition, the new test method can be easily extended to test the linkage heterogeneity in other populations including human beings and it also can be easily applied to other genetic modes in practice.

Extension: testing linkage heterogeneity in an intercross population

Assume the mating type is as follows:

$$\mathrm{P}_1{:}M^1Q^1/m^1q^1 \times \mathrm{P}_2{:}M^2Q^2/m^2q^2,$$

where the superscripts 1 and 2 in M^1 and M^2 denote that the corresponding two alleles inherited from different ancestral sources, respectively; the other notations are similar. The expressions are convenient for us to distinguish the sources of two alleles in the analysis. In this situation, the two parents can produce 16 possible offspring genotypes. The genotypes and the corresponding probabilities are shown in table 1. Under the assumption of recessive inheritance on the trait locus, the observed individuals can be divided into six classes: *mmqq*, *MMqq*, *Mmqq* (diseased individuals); *mmQq*(*QQ*), *MMQq*(*QQ*), *MmQq*(*QQ*) (normal individuals).

Table 1. Offspring genotypes of intercross population and the probabilities.

Gamete2 \ Gamete1	M^1Q^1	m^1q^1	M^1q^1	m^1Q^1
M^2Q^2	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
m^2q^2	$(1-\theta)^2/4$	$(1-\theta)^2/4$	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$
M^2q^2	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$
m^2Q^2	$\theta(1-\theta)/4$	$\theta(1-\theta)/4$	$\theta^2/4$	$\theta^2/4$

Consider F pedigrees, and let N^i denote the number of offspring in the ith pedigree $\left(\sum_{i=1}^{F} N^i = n\right)$ and let r_j^i $(j = 1, ..., 6)$ denote the number of the jth class of offspring in the ith pedigree. For the intercross model, we first need to calculate the MLE of θ via the EM algorithm (Dempster *et al.* 1977). The iterative equation of the recombination rate θ_i in the ith pedigree $(i = 1, 2, \ldots, F)$ is

$$\theta_i^{(t+1)} = \frac{1}{2N^i}\left(2r_2^i + r_3^i + \frac{2}{2-\theta_i^{(t)}}r_4^i + \frac{2\theta_i^{(t)}}{1+\theta_i^{(t)}}r_5^i + \frac{\left[\theta_i^{(t)}\right]^2 + \theta_i^{(t)}}{1-\theta_i^{(t)}+\left[\theta_i^{(t)}\right]^2}r_6^i\right), \quad (2)$$

where $\theta_i^{(t)}$ is the current value of θ_i in the tth iteration. By the form of $\theta_i^{(t+1)}$ in equation (2), we can construct the estimating function

$$g_i(\theta_i) = 2r_2^i + r_3^i + \frac{2}{2-\theta_i}r_4^i + \frac{2\theta_i}{1+\theta_i}r_5^i + \frac{\theta_i^2+\theta_i}{1-\theta_i+\theta_i^2}r_6^i - 2N^i\theta_i.$$

Then, we can get

$$\mathrm{var}\left[g_i(\theta_i)\right] = N^i\left(\frac{\theta_i}{2} - \frac{7\theta_i^2}{2} + \frac{\theta_i}{2-\theta_i} + \frac{(\theta_i^2+\theta_i)^2}{2+2\theta_i^2-2\theta_i} + \frac{\theta_i^2(1-\theta_i)}{1+\theta_i}\right).$$

At the same time, by using data of all the pedigrees, the iteration equation for recombination rate θ under homogeneity via the EM algorithm is

$$\theta^{(t+1)} = \frac{1}{2n}\sum_{i=1}^{F}\left(2r_2^i + r_3^i + \frac{2}{2-\theta^{(t)}}r_4^i + \frac{2\theta^{(t)}}{1+\theta^{(t)}}r_5^i + \frac{\left[\theta^{(t)}\right]^2 + \theta^{(t)}}{1-\theta^{(t)}+\left[\theta^{(t)}\right]^2}r_6^i\right), \quad (3)$$

Table 2. Comparisons of the simulated critical values for the K-test and G-test.

		$F=5$				$F=10$						$F=40$					
		$N=4$		$N=8$		$N=2$		$N=4$		$N=8$		$N=2$		$N=4$		$N=8$	
θ	α	K	G	K	G	K	G	K	G	K	G	K	G	K	G	K	G
0.05	0.01	7.4	8.9	10.7	9.4	11.4	16.1	11.5	18.9	15.8	22.1	33.6	58.9	43.3	63	51.8	62.1
	0.05	6.9	6.3	6.9	8.4	8.6	9.5	11.3	14.8	13.8	15.8	29.8	51.2	38.2	52.6	46.7	51.6
	0.1	4.0	4.2	6.3	5.8	8.6	9.5	11.1	10.3	11.4	14.2	28.1	44.9	35.4	49.2	44.2	49.1
0.1	0.01	10.0	10.6	10.9	12.5	11.7	17.3	17.0	19.4	19.3	22.1	40.9	57.8	51.5	59.4	64.2	62.7
	0.05	7.5	8.9	9.4	9.1	11.4	13.8	14.4	14.9	16.8	16.0	38.1	51.2	47.9	52.1	56.3	52.4
	0.1	6.9	6.3	7.5	6.9	8.9	9.5	12.5	14.0	15.0	13.7	36.3	48.6	45.9	48.9	53.1	48.8
0.2	0.01	10.3	10.7	13.7	12.6	14.2	20	18.3	19.4	22.5	21.2	47.7	57.1	61.5	61.4	67.2	58.8
	0.05	8.9	8.1	10.7	9.3	12.0	15.2	15.9	16.0	18.9	16.9	45.1	51.2	56.6	52.9	60.5	53.6
	0.1	8.0	7.5	8.5	7.7	11.7	13.8	14.8	14.8	16.8	14.8	43.9	48.6	54.0	49.4	57.0	50.2
0.3	0.01	11.4	12.3	13.3	12.5	14.2	17.3	19.4	21.9	21.9	21.3	49.3	55.0	61.1	59.5	63.8	59.0
	0.05	8.9	8.9	10.5	9.3	13.3	15.2	16.1	17.9	18.0	16.9	46.7	49.9	55.5	52.8	57.9	52.9
	0.1	8.9	8.0	8.8	8.0	12.0	13.8	14.8	15.7	15.7	14.7	45.1	48.0	52.7	48.5	53.2	50.2
Critical values of AD*	$\chi^2_{0.01}$	13.3				21.7						62.4					
	$\chi^2_{0.05}$	9.5				16.9						54.6					
	$\chi^2_{0.1}$	7.8				14.7						50.7					

AD*, asymptotic distribution; *F*, pedigree number; *N*, number of offspring in each pedigree.

where $\theta^{(t)}$ is the iteration value of θ in the tth step. By equation (3), the estimate $\hat{\theta}$ of the recombination rate θ can be obtained when the EM algorithm converges. Then the statistic

$$G=\sum_{i=1}^{F}\frac{\left[g_i\left(\hat{\theta}\right)\right]^2}{\mathrm{var}\left[g_i\left(\hat{\theta}\right)\right]},$$

can be constructed and applied to test the linkage heterogeneity for the data of intercross population.

In fact, the recombination rate θ is equal to the probability that a recombinant gamete occurs in linkage analysis. For observed families with T offspring totally ($2T$ gametes), the number of recombinant gametes X obeys the binomial distribution B($2T$, θ). Therefore, once we obtain the number of recombinant gametes in all the observed individuals, we can easily use the proposed G-statistic to test linkage heterogeneity. Besides, the new method is not limited to the assumption of recessive inheritance.

Simulation studies

In this section, simulations are conducted to illustrate and evaluate the performance of our new test method. We respectively compare simulated critical values, type I error rates, and powers of the G-test with those of the existing test methods.

Under the null hypothesis, the asymptotic distributions of the above-mentioned test statistics are determinate. To some degree, the empirical critical values can reflect the closeness of asymptotic distribution of a statistic. In the first simulation, the critical values of each test at levels 0.01, 0.05 and 0.1 are simulated. We also consider three factors that may affect the simulated values: F (the number of pedigrees), N (the number of offspring in each pedigree) and the recombination rate θ. We set the following combinations of F, N and θ: $F=5$, 10, 40; $N=2$, 4, 8; and $\theta=0.05$, 0.1, 0.2, 0.3. Each combination of values is used to generate random sample.

We follow the sampling condition of Risch (1988), i.e., conditional on the total LOD score >0. First, in the simulation, we mainly compare the closeness of the asymptotic

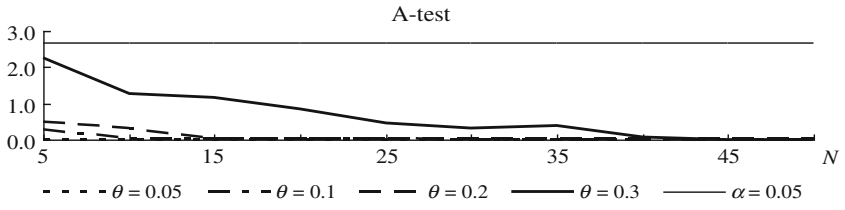

Figure 1. Trend of the critical values of the A-test when N increases ($\alpha=0.05$, $F=2$, and the horizontal line denotes the true critical value of the asymptotic distribution is 2.7).

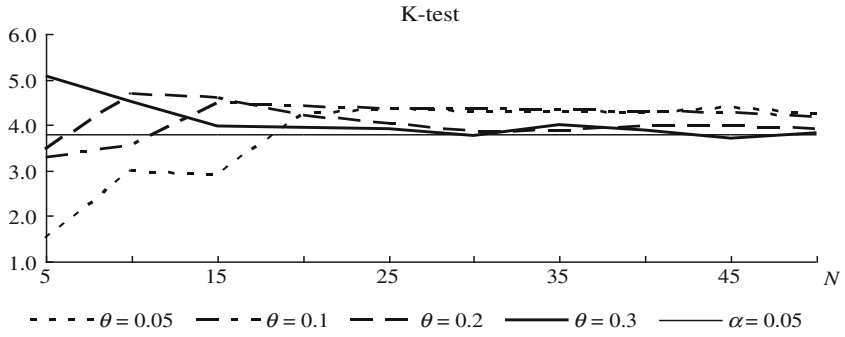

Figure 2. Trend of the critical values of the K-test when N increases ($\alpha = 0.05$, $F = 2$, and the critical value of the asymptotic distribution is 3.8).

distributions of the G-statistic and K-statistic to their own empirical distributions, since they have the same asymptotic distribution. We calculate the values of the statistics from 1000 random samples under each situation of null hypothesis, and then obtain their corresponding empirical critical values (see table 2).

Moreover, we also consider the case of $F = 2$, which is equivalent to dividing the testing families into two subgroups. In this situation, under the condition of total LOD > 0, we simulate the trends of critical values with the varying of family size N for the A-statistic, K-statistic and G-statistic. We let N range from 5 to 50 in increments of 5 and $\theta = 0.05, 0.1, 0.2, 0.3$. The simulation results are listed in figures 1, 2 and 3, respectively.

In the second simulation, we simulate the type I error rates of the K-test, A-test and G-test. We take the significance level $\alpha = 0.05$ and the number of offspring in each pedigree $N = 8$. We also consider the different combination cases of $F = 5, 10, 40$ and $\theta = 0.05, 0.1, 0.2, 0.3$. The simulation results are listed in table 3.

In the third simulation, we consider the comparison of test powers. To compare the simulation results of Risch (1988) and Darlene (1994) with ours, we also apply the same simulation conditions of the alternative hypotheses: take the number of pedigrees $F = 5, 10, 40$ and the pedigree size $N = 8$; set 13 alternative hypotheses that are same with the ones described in Risch (1988) and Darlene (1994) (see table 4). We also take the significance level $\alpha = 0.05$. For one thing, for each mean $\bar{\theta}$ which is calculated by the probability P in table 4, we respectively calculate the values of the K-statistic, A-statistic and G-statistic by 1000 samples satisfying the total LOD > 0 and obtain their own corresponding critical values. We repeat the process 10 times and take the average of the 10 simulated critical values as the threshold of each statistic. For another, under each alternative hypothesis, we randomly generate 1000 samples satisfying the total LOD > 0, calculate the 1000 values for each of the three statistics, respectively, and record the corresponding proportions of 1000 samples with the values beyond the threshold for each test. The simulated powers are presented in table 4.

Evaluation on critical values

From table 2, we can see that most of the simulated values of the K-test are same as that obtained by Risch (1988) and the simulated critical values of our G-statistic are closer to the corresponding thresholds of its asymptotic distribution for each combination of values of F, N and θ. The critical values of the two tests are both impacted by the pedigree number F. However, the increase of F does not seem to improve the closeness of critical values of the K-test, but the critical values of the G-test become closer to the thresholds of its asymptotic distribution. If the critical values of the asymptotic distribution are used to test the linkage heterogeneity, the K-test is likely to provide more type I errors.

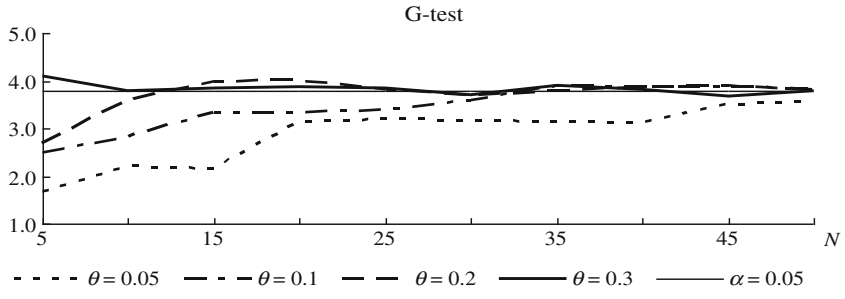

Figure 3. Trend of the critical values of the G-test when N increases ($\alpha = 0.05$, $F = 2$, and the critical value of the asymptotic distribution is 3.8).

Table 3. Type I error rates of the K-test, A-test and G-test.

	$F = 5$			$F = 10$			$F = 40$		
θ	A	K	G	A	K	G	A	K	G
0.05	0.007	0.009	0.007	0.005	0.010	0.035	0.008	0.004	0.041
0.1	0.012	0.034	0.020	0.011	0.042	0.038	0.014	0.091	0.047
0.2	0.018	0.074	0.041	0.020	0.087	0.044	0.018	0.161	0.041
0.3	0.016	0.066	0.041	0.019	0.066	0.042	0.021	0.088	0.041

The significance level is 0.05; $N = 8$.

Figures 1, 2 and 3 respectively illustrate the trend of simulated critical values for the A-, K- and G-test with the varying of N ($F = 2$). We find that the increasing N does not improve the asymptotic effect of the critical values of A-statistic and it will make the critical values smaller. Moreover, smaller the recombination rate, worse the closeness will become. With the increase of N, the critical values of the K-test get closer to, but a little larger than, the thresholds of its corresponding asymptotic distribution. In most cases, the behaviour of the critical values of the G-test is satisfying, i.e. the simulated critical values are stably distributed around the thresholds of its corresponding asymptotic distribution when N increases.

Evaluation of type I errors

The simulation results of type I error rates for the three tests (A-test, K-test and G-test) are listed in table 3. The results indicate that the type I error rates show a trend of increase with the increase of values of F or θ for all the three tests. The A-test is very conservative and the largest value of the type I error rates of the A-test is only 0.021 (the nominal level is 0.05). The conservativeness and nonconservativeness situations of K-test exist together. The type I error rates of G-test are a little better compared with the ones of the other tests.

Evaluation on powers

From table 4, we can see that the majority of simulation results of powers for the K- and A-tests are closed to the results of Risch (1988) or Darlene (1994). The small difference may be due to the slight difference of the used critical values for two tests.

The powers of the three tests are all affected by the distribution of θ, e.g., increasing the probability of $\theta = 0.5$ will improve the test powers in the same type of linkage heterogeneity; the powers in cases 1, 2, 3 and 4 are evidently higher than the ones in cases 5 and 6, i.e., the heterogeneity test between pedigrees of close linkage and nonlinkage is more powerful. And as expected, the powers of the three methods are all lower, when the value of F is smaller and the values all increase with the increase of F. From the simulation results under the alternative hypotheses of the last seven cases (cases 7–13), for each test method, the different combinations of linkage degrees (close linkage, $\theta \leq 0.1$; medium

Table 4. Power comparisons of the K-test, A-test and G-test.

Case	θ				P		$\bar{\theta}$	$F = 5$			$F = 10$			$F = 40$		
								A	K	G	A	K	G	A	K	G
1	0		0.50	0.9		0.1	0.05	0.35	0.35	0.38	0.59	0.59	0.64	0.98	0.85	0.99
2	0		0.50	0.8		0.2	0.1	0.62	0.62	0.62	0.85	0.78	0.87	1	0.99	1
3	0.05		0.50	0.8		0.2	0.14	0.53	0.48	0.50	0.78	0.7	0.77	1	0.99	1
4	0.05		0.50	0.5		0.5	0.275	0.67	0.65	0.65	0.92	0.92	0.93	1	1	1
5	0.20		0.50	0.8		0.2	0.26	0.19	0.17	0.20	0.30	0.24	0.30	0.70	0.58	0.70
6	0.20		0.50	0.5		0.5	0.35	0.23	0.21	0.25	0.37	0.36	0.38	0.87	0.82	0.86
7	0		0.10	0.5		0.5	0.05	0.07	0.10	0.17	0.13	0.17	0.18	0.28	0.40	0.43
8	0.05		0.15	0.5		0.5	0.1	0.10	0.10	0.09	0.11	0.12	0.13	0.14	0.22	0.23
9	0.10		0.30	0.25		0.75	0.25	0.11	0.12	0.11	0.18	0.19	0.17	0.29	0.38	0.36
10	0.20		0.40	0.5		0.5	0.3	0.14	0.13	0.15	0.19	0.18	0.19	0.41	0.43	0.42
11	0.05	0.15	0.25	0.25	0.5	0.25	0.15	0.12	0.12	0.12	0.15	0.18	0.16	0.23	0.33	0.33
12	0.10	0.20	0.30	0.25	0.5	0.25	0.2	0.11	0.10	0.11	0.11	0.12	0.12	0.23	0.25	0.24
13	0.05	0.20	0.35	0.25	0.5	0.25	0.2	0.18	0.19	0.18	0.24	0.28	0.27	0.52	0.68	0.64

Power calculated for the 0.05 significance level.

linkage, $0.1 < \theta \leq 0.2$; loose linkage, $\theta > 0.2$) also mean different test powers.

In all the simulated cases, most of the time the power of the G-test is higher, although the difference is not large. However, in cases 7–13 ($\theta \neq 0.5$ in each pedigree), the power status of the A-test is generally poor. In practice, the used critical value of the A-test is unusually obtained from the grid search method (Hauser *et al.* 2004), which is smaller than the true one in general, so the factual power of the A-test will be lower than the simulated power.

In addition, we considered another simulation to compare the computing time of performing K-test, A-test and G-test. For the same simulation data that we generated in the previous power comparison, we recorded the average computing time for performing each test. We found that the computing speed of performing G-test is much faster than those of performing A-test and K-test. Taking case 5 in table 4 for example, when $F = 5$, the average computing time is ∼8 s for performing G-test, but it is ∼1 min for K-test and ∼5 min for A-test, respectively (the computing time all contains the time of obtaining simulated critical values). This is expected because the computing amount of G-test is less, i.e., the G-test only needs estimating recombination fraction θ under H_0, however, the other test methods need estimating all parameters under both H_0 and H_1, which is much time-consuming.

Real example

Gene probe is an important tool for gene mapping in the genetic linkage analysis. The accuracy of the location of gene probe is essential to detect disease loci. The probe ST14 are located at the end of X chromosome and it has been recommended for the research of three common genetic diseases: haemophilia A, mental retardation with a fragile X chromosome and adrenal leukodystrophy (Oberlé *et al.* 1985; Laird 1987).

Applying the G-test proposed in this paper, we will test the linkage heterogeneity for the locus ST14 and another locus F9 located on the X chromosome. In fact, loci F9 and ST14 are genetically linked with an overall recombination distance of 0.28 (±0.059) in normal populations (Drayna and White 1985). However, in the pedigrees with fragile X syndrome, there is linkage heterogeneity between the two loci.

We randomly sampled eight pedigrees and their 70 offspring from the pedigree data provided in Oberlé *et al.* (1986) and Brown *et al.* (1987), where five pedigrees (no. 1, 2, 5, 9 and 13) came from Oberlé *et al.* (1986) and three pedigrees (no. 20, 22 and 52) came from Brown *et al.* (1987). Table 5 illustrates the data in detail. In pedigree 1, the genotype data of offspring 7, 8, 10 and 11 in generation IV are randomly sampled to use. The explanation for data of other pedigrees is completely similar. In the 70 offspring of the eight considered pedigrees, there are totally 23 recombinants.

After calculating by the proposed method, we obtain the value of statistic G is 19.8793 and the corresponding P value = 0.0058. It is concluded that in the eight pedigrees, the linkage heterogeneity exists between loci F9 and ST14 on X chromosome. This conclusion is the same as the one given in Brown *et al.* (1987). Through the illustration of this example, we hope that the test problem of linkage heterogeneity can attract more attention and the relevant results obtained from the test can provide us valuable references on the detection of disease loci. In a word, we can not ignore the key step in practical linkage analysis.

Discussion

In this paper, we propose a new test method, called G-test, the theoretical basis of which is not the likelihood-ratio and the Bayesian theory. The method applies to both experimental (e.g., BC and F_2) populations and human population. The calculation of the G-statistic is easier than those of the B-test and D-test. Once the number of recombinant gametes of offspring is determined in the tested pedigrees, the value of this statistic can be easily calculated without complicated parameter estimation. At the same time, we can see that the problem solved by the G-test and K-test are same, but the results exist difference in varying degrees.

Table 5. Data information including eight different pedigrees in the example.

Code of pedigree	Code of offspring	Number of offspring	Number of recombinants
1	IV* (7, 8, 10, 11)	4	0
2	III (3 ∼ 5)	3	0
5	III (1 ∼ 6)	6	3
9	II (1 ∼ 8)	8	2
13	III (3 ∼ 15, 17, 18)	15	2
20	IV (1 ∼ 7, 9, 10, 18, 20 ∼ 24, 32, 33)	17	7
22	IV (4 ∼ 10)	7	1
52	III (1, 3, 5, 6, 8, 10, 12 ∼ 15)	10	8
Total	—	70	23

IV* denotes generation IV.

The proposed G-test has its own advantages: the closeness degree of the asymptotic distribution of the G-statistic under the null hypothesis is better than that of the K-statistic; the conservativeness of the G-test is lower than that of the A-test, and the K-test seems very sensitive to θ from the results in table 3, which is consistent with that provided in Risch (1988). In our simulation, although we only listed the results on comparing the power of the new test with that of the A-test and K-test, in fact we also compared the performance of G-test with that of B-test and D-test by contrasting the historical simulation data under the same situations in Darlene (1994). We reach the same conclusion that the power of G-test is higher in most cases, i.e., the G-test has the highest power in 30 of the total 39 simulation cases (13 alternative hypotheses, and $F = 5$, 10 and 40) and the advantage is more apparent when the number of pedigrees is smaller ($F = 5$, 10).

All in all, from our simulation results and experience, we suggest that it is reasonable to take advantage of the G-test and the corresponding critical values of its asymptotic distribution to test linkage heterogeneity when $F > 5$ and $N > 4$, to avoid unnecessary inference error; if the values of F and N are all not large in practice, one can apply the simulated critical values presented in table 2 when performing the G-test.

The last but not the least, for the test of linkage heterogeneity, we generally talk about the situation in separate pedigrees, where each pedigree can be seen as a subgroup, because the observational data of pure big pedigree is difficult to collect in practice. According to the actual demand, some indexes can be considered, such as geographical origin, ancestry, gender, age, etc., for purpose of dividing subgroups to investigate the problem of linkage heterogeneity. Thus, by testing the data with a smaller number of subgroups and a large number of individuals within each subgroup, the power of testing linkage heterogeneity can be improved and more reliable results can be obtained by our experience. Further, if there exists linkage heterogeneity in the tested subgroups, then we can make a new division for heterogeneity test within the subgroup to find new difference.

Appendix

Lemma. Consider a univariate parameter θ. Let x_i represent a random sample of size n_i from an underlying distribution in the ith experiment. We use an estimating function for θ given x_i which is denoted by $g_i(x_i;\theta)$. Assume that each estimating function satisfies the assumptions below:

(1) $E\{g_i(x_i;\theta)\} = 0$, for any θ,
(2) $E\left\{\frac{\partial}{\partial\theta} g_i(x_i;\theta)\right\} < 0$, for any θ,
(3) There exists a variance function $V_i(\theta)$ such that

$$\frac{g_i(x_i;\theta)}{\sqrt{V_i(\theta)}} \to N(0,1) \quad \text{as} \quad n_i \to \infty.$$

Let the estimator $\hat{\theta}$ for the parameter θ be defined by the solution of the equation $\left\{\sum_{i=1}^{F} g_i(x_i;\theta)\right\} = 0$. Consider the statistic $G = \sum_{i=1}^{F} \frac{\left[g_i\left(\hat{\theta}\right)\right]^2}{\mathrm{var}\left[g_i\left(\hat{\theta}\right)\right]}$. Let D_1 and D_2 be independent random variables that follow chi-squared distributions with $F-2$ and 1 degree of freedom, respectively. Then the statistic G is asymptotically equivalent to $D_1 + cD_2$ in the large sample cases, where

$$c = \frac{\sum_{j=1}^{F} V_j(\theta) \sum_{i=1}^{F} \left\{E\left[\frac{\partial}{\partial\theta} g_i(x_i;\theta)\right]\right\}^2 / V_i(\theta)}{\left[\sum_{i=1}^{F} E\left\{\frac{\partial}{\partial\theta} g_i(x_i;\theta)\right\}\right]^2}.$$

See the Lemma and its proof in Fujii (1994).

Acknowledgements

This research was supported by the National Natural Science Foundation of China (nos. 11201129 and 11371083), the Natural Science Foundation of Heilongjiang Province of China (A201207), the Scientific Research Foundation of Department of Education of Heilongjiang Province of China (nos. 1253G044 and 12531508) and the Scientific Foundation of Heilongjiang University for Distinguished Young Scholars (no. JCL 201003).

References

Bisceglia L., Zoccolella S., Torraco A., Piemontese M. R., Dell'Aglio R., Amati A. *et al.* 2010 A new locus on 3p23–p25 for an autosomal-dominant limb-girdle muscular dystrophy, LGMD1H. *Eur. J. Hum. Genet.* **18**, 636–641.

Brown W. T., Jenkins E. C., Gross A. C., Chan C. B., Krawczun M. S., Duncan C. J. *et al.* 1987 Further evidence for genetic heterogeneity in the fragile X syndrome. *Hum. Genet.* **75**, 311–321.

Burmeiste M., McInnis M. G. and Zöllner S. 2008 Psychiatric genetics: progress amid controversy. *Nat. Rev. Genet.* **9**, 527–540.

Buxbaum J. D., Silverman J. M., Smith C. J., Kilifarski M., Reichert J., Hollander E. *et al.* 2001 Evidence for a susceptibility gene for autism on chromosome 2 and for genetic heterogeneity. *Am. J. Hum. Genet.* **68**, 1514–1520.

Laird C. D. 1987 Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation. *Genetics* **117**, 587–599.

Darlene R. 1994. A combined test of linkage heterogeneity. *Am. J. Hum. Genet.* **55**, 841–848.

Dempster A. P., Laird N. M. and Rubin D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38.

Drayna D. and White R. 1985 The genetic linkage map of the human X chromosome. *Science* **230**, 753–758.

Fujii Y. 1994 On homogeneity test using estimating function. *Bull. Informat. Cybern.* **26**, 101–107.

Grice D. E., Halmi K. A., Fichter M. M., Strober M., Woodside D. B., Treasure J. T. *et al.* 2002 Evidence for a susceptibility gene for anorexia nervosa on chromosome 1. *Am. J. Hum. Genet.* **70**, 787–792.

Hauser E. R., Richard M., Duren W. L., Bass M. P., Langefeld C. D. and Boehnke M. 2004 Ordered subset analysis in genetic linkage mapping of complex traits. *Genet. Epidemiol.* **27**, 53–63.

Hodge S. E., Anderson C. E., Neiswanger K., Sparkes R. S. and Rimoin D. L. 1983 The search for heterogeneity in insulin-dependent diabetes mellitus (IDDM): linkage studies, two-locus models, and genetic heterogeneity. *Am. J. Hum. Genet.* **35**, 1139–1155.

Huber C., Delezoide A. L., Guimiot F., Baumann C., Malan V., Le Merrer M. *et al.* 2009 A large-scale mutation search reveals genetic heterogeneity in 3M syndrome. *Eur. J. Hum. Genet.* **17**, 395–400.

Leal S. M. and Ott J. 2000 Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. *Am. J. Hum. Genet.* **66**, 567–575.

Lewis T. B., Leach R. J., Ward K., O'Connell P. and Ryan S. G. 1993 Genetic heterogeneity in benign familial neonatal convulsions: identification of a new locus on chromosome 8q. *Am. J. Hum. Genet.* **53**, 670–675.

McClellan J. and King M. C. 2010 Genetic heterogeneity in human disease. *Cell* **141**, 210–217.

Morton N. E. 1956 The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am. J. Hum. Genet.* **8**, 80–96.

Oberlé I., Drayna D., Camerino G., White R. and Mandel J. L. 1985 The telomeric region of the human X chromosome long arm: presence of a highly polymorphic DNA marker and analysis of recombination frequency. *Proc. Natl. Acad. Sci. USA* **82**, 2824–2828.

Oberlé I., Heilig R., Moisan J. P., Kloepfer C., Mattéi G. M., Mattéi J. F. *et al.* 1986 Genetic analysis of the fragile-X mental retardation syndrome with two flanking polymorphic DNA markers. *Proc. Natl. Acad. Sci. USA* **83**, 1016–1020.

Ott J. 1977 Counting methods (EM algorithm) in human pedigree analysis: linkage and segregation analysis. *Ann. Hum. Genet.* **40**, 443–454.

Ott J. 1999 *Analysis of human genetic linkage*, 3rd edition. Johns Hopkins University Press, Baltimore, USA.

Risch N. 1988 A new statistical test for linkage heterogeneity. *Am. J. Hum. Genet.* **42**, 353–364.

Risch N. and Baron M. 1982 X-linkage and genetic heterogeneity in bipolar related major affective illness: reanalysis of linkage data. *Ann. Hum. Genet.* **46**, 153–166.

Smith C. A. B. 1963 Testing for heterogeneity of recombination values in human genetics. *Ann. Hum. Genet.* **27**, 175–182.

Stine O. C., Xu J., Koskela R., McMahon F. J., Gschwend M., Friddle C. *et al.* 1995 Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. *Am. J. Hum. Genet.* **57**, 1384–1394.

Received 3 June 2012, in revised form 28 February 2013; accepted 3 June 2013
Published on the Web: 13 November 2013