

The Power of Transmission Disequilibrium Tests for Quantitative Traits

Jinming Li[†], Dai Wang[†], Jianping Dong, Renfang Jiang, Kui Zhang, Shuanglin Zhang, Hongyu Zhao, Fengzhu Sun^{*}

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, Connecticut (J.L., K.Z., S.Z., H.Z.), Department of Mathematics, University of Southern California, Los Angeles, California (D.W., K.Z., F.S.), Department of Mathematical Sciences, Michigan Technological University, Houghton, Michigan (J.D., R.J.)

We develop a score statistic to test for linkage in the presence of linkage disequilibrium for quantitative traits. We then extend this method to analyze multiple tightly linked markers. One potential limitation with the use of many genetic markers is the large number of degrees of freedom involved that may reduce the overall power to detect linkage. To overcome this limitation, we propose to group haplotypes on the basis of haplotype similarity before performing transmission disequilibrium tests. Finally, we apply these methods to the GAW12 simulated data and compare their power.

Key words: transmission/disequilibrium test, quantitative trait, haplotype

INTRODUCTION

The transmission disequilibrium test (TDT) for linkage introduced by Spielman et al. [1993] employs a family-based design and is robust to population stratification. Schaid [1996] proposed a general score statistic for the TDT. In this article, we extend Schaid's results and develop a score statistic for quantitative traits. We show that the tests for

[†] These authors contributed equally to this work.

^{*} Correspondence to: Dr. Fengzhu Sun, Department of Mathematics, University of Southern California, Los Angeles, CA 90089. Fax: 213-740-2424. E-mail: fsun@hto.usc.edu

Li et al.

quantitative traits of Rabinowitz [1997], Lunetta et al. [2000], and Sun et al. [2000] are special cases of our general score statistic.

With the development of single nucleotide polymorphic markers (SNP), dense genetic maps will be available within a candidate region. Zhao et al. [2000] proposed a method to simultaneously analyze multiple tightly linked markers for dichotomous traits. We extend the idea of Zhao et al. [2000] and propose a score statistic for quantitative trait using multiple tightly linked markers. To reduce the degree of complexity among the multisite haplotypes and to increase the overall statistical power to detect linkage, we also perform the TDTs after haplotypes are grouped on the basis of haplotype similarity. We apply these methods to the simulated data of Genetic Analysis Workshop 12 and compare the power of these test statistics under different circumstances.

METHODS

Score statistics

Suppose we have N offspring with their parents' genotypes available. Let Q_i denote the trait value of the i th offspring, and let g_i , $g_{i,m}$, and $g_{i,f}$ denote the marker genotypes of the offspring, the mother, and the father, respectively. Let $f(Q_i | g_i, g_{i,m}, g_{i,f})$ be the distribution of the offspring's trait value conditional on the genotypes of the offspring and the parents. Using Bayes rule, the probability of the genotype of the offspring conditional on both the offspring's trait value and the parents' marker genotypes can be written as

$$P(g_i | g_{i,m}, g_{i,f}, Q_i) = \frac{f(Q_i | g_i, g_{i,m}, g_{i,f})P(g_i | g_{i,m}, g_{i,f})P(g_{i,m}, g_{i,f})}{\sum_{g_i^* \in G_i} f(Q_i | g_i^*, g_{i,m}, g_{i,f})P(g_i^* | g_{i,m}, g_{i,f})P(g_{i,m}, g_{i,f})}, \quad (1)$$

where the summation in the denominator is over the four possible marker genotypes that the parents can produce and g_i^* is one of them.

We assume that, given the marker genotype of the offspring, the distribution of the trait value of the offspring does not depend on the marker genotypes of the parents. We also assume that the trait value has a normal distribution with fixed variance. A convenient approach is to assume that the probability density function of the trait value of an offspring with genotype g_i is

$$f_{g_i}(Q_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(Q_i - c - X_i'\beta)^2}{(2\sigma^2)}\right\},$$

where X_i is the code of the i th offspring's genotype g_i . In this paper, an additive model is used, that is, the k th element of X_i is the number of copies of the k th allele in the genotype g_i . Ignoring constants, the likelihood function is

$$L = \prod_{i=1}^N \frac{\exp\left\{-\frac{(Q_i - c - X_i'\beta)^2}{(2\sigma^2)}\right\}}{\sum_{g_i^* \in G_i} \exp\left\{-\frac{(Q_i - c - X_i^{*'}\beta)^2}{(2\sigma^2)}\right\}}.$$

There are several methods to derive the test statistic. One approach is Rao's score statistic. The general form of the score statistic is

$$S = UV^{-1}U,$$

where $U = \partial \ln L / \partial \beta |_{\beta=0}$ and V is the Fisher's information matrix with elements

$$V_{ij} = -E\left[\partial^2 \ln L / \partial \beta_i \partial \beta_j |_{\beta=0}\right].$$

Under our assumptions, both U and V can be calculated.

Let X_{ij} be the vector for the j th genotype g_j in the set G_i , and let $\bar{X}_i = \frac{1}{4} \sum_{j=1}^4 X_{ij}$.

Also, let $V_i = \frac{1}{4} (\sum_{j=1}^4 X_{ij} X_{ij}') - \bar{X}_i \bar{X}_i'$ be the covariance matrix for the four X_{ij} vectors.

Then, under the normal distribution assumption, U and V can be written as

$$U = \sum_{i=1}^N \frac{Q_i - c}{\sigma^2} (X_i - \bar{X}_i), \quad V = \sum_{i=1}^N \frac{(Q_i - c)^2}{\sigma^4} V_i.$$

Under the null hypothesis of no linkage, S has an approximate chi-square distribution with the number of degrees of freedom equals the rank of V when the sample size is large. The constant c is the population mean of the trait value. The above test is valid for an arbitrary constant c , but the power of these test statistics depends on the choice of c [Lunetta et al. 2000; Sun et al., 2000]. Replacing c with the average trait value, \bar{Q} , of all the offspring in the sample, the above score statistic is the same as that of Rabinowitz [1997].

Another approach is to obtain the maximum of the χ^2 statistic associated with each allele. Let $S_l = U_l^2 / V_{ll}$. S_l is the test statistic of the l th allele vs. all other alleles. Let

$S_{\max} = \max_{l=1}^k S_l$, where k is the number of alleles at the marker locus. The distribution of S_{\max} can be estimated by simulation. For each simulated sample, the two alleles in the parents are randomly transmitted to each offspring. The test statistics S_{\max} from the simulated data sets are then used as the reference distribution to assess the significance level of the observed test statistic.

Extension to haplotypes

Both S and S_{\max} introduced above analyze one marker at a time. With the rapid progress of the Human Genome Project, many polymorphic genetic markers can now be identified and genotyped within a very small region of a chromosome. Studying multiple markers jointly may yield more information for linkage. In an effort to fully utilize multiple tightly linked markers, Zhao et al. [2000] proposed a method to test for linkage for dichotomous traits with multiple tightly linked markers. We extend the method to quantitative traits. The method can be briefly described as follows.

Let h be the total number of possible haplotypes and $\{p_j, j=1, \dots, h\}$ be an arbitrary set of haplotype frequencies. We will use j to denote the j th haplotype. Suppose $G_i = \{g_i, g_{i,m}, g_{i,f}\}$ is the i th family's *genotype group*, where $g_i, g_{i,m}$, and $g_{i,f}$ are the genotypes of the offspring, the mother, and the father, respectively. Let (j,l,km) denote the event that the father has haplotype pair (j,k) and transmits j to the offspring, and the mother has haplotype pair (l,m) and transmits l to the offspring. The event (j,l,km) is referred to as compatible with genotype group G_i if the resulting genotypes are the same as in G_i . (j,l,km) will be called a *haplotype group*. See Zhao et al. [2000] for examples on these definitions. If event (j,l,km) is compatible with genotype group G_i , then the probability of (j,l,km) conditional on G_i is

$$w_{G_i}^{(j,l,km)} = \frac{p_j p_k p_l p_m}{\sum_{(j^* l^*, k^* m^*) \in G_i} p_{j^*} p_{k^*} p_{l^*} p_{m^*}}.$$

The summation is over all the haplotype groups compatible with the genotype group G_i . Let $X_i(j,l)$ be any reasonable coding vector of haplotype pair (j,l) . In this paper,

Li et al.

we assume an additive model. Let $\bar{X}_i(jl, km)$ and $V_i(jl, km)$ be the mean and covariance of the coded vectors of all the possible haplotype pairs in a family with parents' haplotype pairs (j, k) and (l, m) , respectively. Then,

$$U = \sum_{i=1}^N \frac{Q_i - c}{\sigma} \sum_{(jl, km) \in G_i} w^{(jl, km)}_{G_i} (X_i(j, l) - \bar{X}_i(jl, km))$$

$$V = \sum_{i=1}^N \frac{(Q_i - c)^2}{\sigma^2} \sum_{(jl, km) \in G_i} w^{(jl, km)}_{G_i} V_i(jl, km)$$

Corresponding to S and S_{\max} , we can obtain two new statistics. We denote these two new statistics by T and T_{\max} . The distribution of T and T_{\max} can be estimated by simulation.

Zhao et al. [2000] discussed three methods to determine haplotype frequencies $\{p_j, j=1, \dots, h\}$. Here we adopt the EM algorithm to estimate the p_j . Assume that all parents are a random sample of unrelated individuals from a population with Hardy-Weinberg equilibrium. Suppose n is the total number of parents and n_g is the number of parents with genotype g . The EM algorithm starts by assigning all haplotype frequencies to be equal. That is, $p_{0j} = 1/h$. Let $n_{m,jk}$ be the estimated number of parents with haplotype pair (j, k) . Then the iterative formulas of the EM algorithm are

$$n_{m,jk} = n_g \frac{p_{m,j} p_{m,k}}{\sum_{(j^*, k^*) \in g} p_{m,j^*} p_{m,k^*}}, \quad p_{m+1,j} = \frac{2n_{m,jj} + \sum_{k \neq j} n_{m,jk}}{2n}$$

The iterations are stopped when the Euclidean distance between $\{p_{m,j}, j=1, \dots, h\}$ and $\{p_{m+1,j}, j=1, \dots, h\}$ is smaller than a sufficiently small value, e.g., 10^{-6} .

Haplotype grouping

When the number of haplotypes in the study is very large, there are too many degrees of freedom in the statistical tests. This may lead to decreased power to detect linkage. To reduce the complexity among many haplotypes, we propose to group haplotypes on the basis of haplotype similarity. In our analysis, we use the `hclust()` and `cutree()` in `Splus` to divide all haplotypes into two groups using the hierarchical clustering method. For SNPs, the dissimilarity between two haplotypes is simply defined as the number of different bases between the two haplotypes. After haplotype grouping, we can consider the two groups as two different alleles in our analysis. For this "two-allele" system, we use the method studied by Sun et al. [2000] to detect linkage. More specifically, let l_j denote the number of the offspring in the j th family, and let $Q_{j,k}$ denote the trait value of the k th offspring in the j th family, where $k = 1, \dots, l_j$. Suppose all haplotypes are divided into two groups, M and N . Define $Y_{j,k}^{(m)} = 1$ (or -1) if the mother in the j th family is heterozygous (with respect to M and N) and transmits the M (or N) to the k th offspring, and $Y_{j,k}^{(m)} = 0$ if the mother is homozygous (with respect to M and N). We similarly define $Y_{j,k}^{(f)}$ for the father. Let \bar{Q} denote the mean trait value of all children in all of the families. Conditional on the trait values and the parental genotypes, the statistic

$$s_j = \sum_{k=1}^{l_j} (Q_{j,k} - \bar{Q})(Y_{j,k}^{(f)} + Y_{j,k}^{(m)})$$

has mean 0 under the null hypothesis of no linkage, and the conditional variance of s_j can be estimated by

$$\sigma_j^2 = \sum_{k=1}^{l_j} (Q_{j,k} - \bar{Q})(|Y_{j,k}^{(f)}| + |Y_{j,k}^{(m)}|).$$

The test statistic for quantitative traits is then $T = \sum s_j / \sqrt{\sum \sigma_j^2}$.

RESULTS AND DISCUSSION

We apply the above methods to the GAW12 data. According to the answers distributed by GAW12, we know that D19G032 and D02G170 are the closest markers to the QTL of trait 1 and trait 2, D09G118 is closest to the QTL of trait 2 and trait 3, and D01G139 is closest to the QTL of trait 5. We concentrate on the regions around those genetic loci and consider those loci and the closest 0, 1, 2, or 3 loci around them on each side. Thus the number of loci being considered, L , is 1, 3, 5, or 7.

For both the general and the isolated populations, we calculate the test statistics for each replicate and then obtain the approximate power from the 50 replicates. The significance level α is set to be 0.05. To illustrate that the procedure yields appropriate false-positive rates, we include a marker, D01G004, which is almost unlinked (with a recombination fraction of 0.46) to the only QTL on chromosome 1. In all the situations considered, we let $c = \bar{Q}$, where \bar{Q} is the mean trait value of the offspring. We first consider the loci under study separately and then combine the results using the Bonferroni correction. If the smallest p -value for all the loci is less than $0.05/L$, we say the signal is detected. Table I lists the power of S and S_{\max} for different values of L in the general and the isolated populations.

TABLE I. The power of S and S_{\max} with $c = \bar{Q}$ for the general and the isolate populations

		General Population				Isolate Population			
		$L=1$	$L=3$	$L=5$	$L=7$	$L=1$	$L=3$	$L=5$	$L=7$
D19G032	S	0.06	0.00	0.02	0.02	0.06	0.12	0.14	0.14
Trait 1	S_{\max}	0.06	0.00	0.04	0.06	0.06	0.14	0.20	0.16
D19G032	S	0.08	0.04	0.02	0.02	0.06	0.04	0.06	0.04
Trait 2	S_{\max}	0.06	0.04	0.06	0.02	0.06	0.08	0.10	0.12
D02G170	S	0.04	0.06	0.08	0.12	0.10	0.24	0.16	0.16
Trait 1	S_{\max}	0.08	0.10	0.16	0.20	0.10	0.16	0.20	0.18
D02G170	S	0.06	0.10	0.10	0.06	0.14	0.22	0.30	0.26
Trait 2	S_{\max}	0.08	0.20	0.10	0.14	0.18	0.20	0.22	0.18
D09G118	S	0.10	0.08	0.06	0.10	0.24	0.20	0.10	0.12
Trait 2	S_{\max}	0.10	0.06	0.04	0.08	0.22	0.16	0.16	0.12
D09G118	S	0.10	0.22	0.14	0.10	0.30	0.48	0.32	0.34
Trait 3	S_{\max}	0.12	0.22	0.20	0.18	0.30	0.40	0.32	0.28
D01G139	S	0.08	0.24	0.16	0.22	0.46	0.66	0.64	0.64
Trait 5	S_{\max}	0.06	0.16	0.12	0.12	0.50	0.54	0.48	0.44
D01G004	S	0.00	0.04	0.04	0.06	0.02	0.04	0.02	0.04
Trait 5	S_{\max}	0.02	0.02	0.02	0.02	0.00	0.02	0.02	0.02

For the GAW12 data, the power of S and S_{\max} is relatively low, usually below 30% for most of the loci except D09G118 for trait 3 and D01G139 for trait 5 using the isolated population. From Table I, we can see that: (a) For most of the situations, the sampling variability is sufficiently large that no definite conclusions can be drawn for the power of S and S_{\max} . For the last four rows using the isolated population when the power is relative high, S seems to be more powerful than S_{\max} . (b) For all the loci, the power of the tests using the isolated population is usually higher than the power of the tests using the general population. This is reasonable because the isolated population has a shorter

history and usually keeps stronger association than the general population. (c) Considering only one marker is not as powerful as considering multiple markers together ($L > 1$), while considering too long regions around the genetic loci may not increase the power of the test.

To examine the benefit of haplotype grouping before conducting the transmission disequilibrium tests, we analyze the sequence data from the seven candidate genes. All of the 50 replications for the isolated population are analyzed. For each replication, the data consist of 23 pedigrees with 1497 individuals. Because only genotype data are observed for people in the pedigrees, haplotype compositions for a given individual have to be inferred from the pedigree data. In our analysis, we use the haplotype reconstruction program HAPLORE (unpublished results) to reconstruct haplotypes from genotype data. Even with the help of many relatives in an extended pedigree, haplotypes cannot always be uniquely identified. Among the 23 pedigrees in each replication, we only select nuclear families in which the genotypes are available from both parents. For different genes and different replicates, the number of nuclear families that can be selected varies from 40 to 160, and the number of different haplotypes ranges from 30 to 150. We set the statistical significance level at 0.05.

For the five quantitative traits, there is almost no power to detect linkage before haplotype grouping (results not shown). This is in part due to the large number of haplotypes examined. In Table II, we summarize the power to detect linkage between candidate genes and quantitative traits after haplotype grouping. The power is estimated from 50 replicated populations. It is apparent that haplotype grouping substantially increases the power to detect linkage, even for genes that carry multiple functional alleles.

TABLE II. The power of the tests to detect linkage between candidate genes and quantitative traits from 50 replicated samples. All haplotypes are divided into two groups using the hierarchical clustering method. Three candidate genes with chromosomal designations are labeled with their chromosomal numbers. There are three true gene-trait associations among the 35 pairs. The power to detect these three true associations are bold faced in the table.

	Power				
	Q_1	Q_2	Q_3	Q_4	Q_5
Gene 1 (Chr. 6)	0.08	0.14	0.06	0.08	0.14
Gene 2 (Chr. 1)	0.02	0.04	0.04	0.02	0.54
Gene 3	0.02	0.00	0.08	0.06	0.02
Gene 4	0.00	0.10	0.00	0.02	0.00
Gene 5	0.00	0.02	0.04	0.04	0.08
Gene 6 (Chr. 19)	0.98	0.80	0.06	0.08	0.04
Gene 7	0.02	0.06	0.02	0.06	0.02

ACKNOWLEDGEMENTS

Supported in part by DK53392 to F.S. and GM59507 and HD36834 to H.Z. from NIH.

REFERENCES

Lunetta KL, Faraone SV, Biederman J, Laird NM (2000): Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am J Hum Genet* 66:605-614

Rabinowitz DA (1997): A transmission disequilibrium test for quantitative trait loci. *Hum Hered* 47:342-350.

Schaid DJ (1996): General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 13:423-449.

Spielman RS, McGinnis RE, Ewens WJ (1993): Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506-516.

Sun FZ, Flanders WD, Yang QH, Zhao HY (2000): Transmission disequilibrium tests for quantitative traits. *Ann Hum Genet*. In Press.

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun FZ, Kidd KK (2000): Transmission /disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 67:936-946.