

Test of Association for Quantitative Traits in General Pedigrees: The Quantitative Pedigree Disequilibrium Test

Shuanglin Zhang, Kui Zhang, Jinming Li, Fengzhu Sun, Hongyu Zhao*

Department of Epidemiology and Public Health (S.Z., K.Z., J.L., H.Z), Yale University School of Medicine, New Haven, Connecticut; Department of Mathematics (F.S.), University of Southern California, Los Angeles, California

Many statistical methods have been proposed in recent years to test for genetic linkage and association between genetic markers and traits of interest through unrelated nuclear families. However, most of these methods are not valid tests of association in the presence of linkage when some of the nuclear families are related. As a result, related nuclear families in large pedigrees cannot be included in a single analysis to test for linkage disequilibrium. Recently, Martin et al. [2000] proposed the Pedigree Disequilibrium Test (PDT) to test for linkage and association in general pedigrees for qualitative traits. In this article, we develop a similar Quantitative Pedigree Disequilibrium Test (QPDT) to test for linkage and association in general pedigrees for quantitative traits. We apply both the PDT and the QPDT to analyze the sequence data from the seven candidate genes in the simulated data sets in the 12th Genetic Analysis Workshop.

Key words: transmission/disequilibrium test, quantitative trait, qualitative trait, general pedigrees

INTRODUCTION

The transmission/disequilibrium test (TDT) [Spielman et al., 1993] may offer a powerful alternative approach in the identification of genes related to complex traits. Although the original TDT proposed by Spielman et al. [1993] was only applicable to qualitative traits with genotype information from the affected offspring and both of its parents, much research has been done in recent years to extend the TDT to study both qualitative and quantitative traits using families with different family structures. See Zhao [2000] for an updated review on various extensions of the original TDT. Despite these

* Correspondence to: Dr. Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520. Fax: 203-785-6912. E-mail: hongyu.zhao@yale.edu

advantages, one limitation of most of these methods is that, when related nuclear families from extended pedigrees are analyzed together, although these methods remain valid tests of linkage, they are not valid tests of association. Therefore, it is desirable to have a valid test of linkage disequilibrium that can use all potentially informative data from extended pedigrees. Abecasis et al. [2000] developed a variance components approach to detecting association for a nuclear family with an arbitrary number of sibs, with or without parents. Martin et al. [2000] proposed the Pedigree Disequilibrium Test (PDT) to detect linkage disequilibrium between a genetic marker and qualitative traits in general pedigrees. To generalize the PDT to quantitative traits, we have developed the Quantitative Pedigree Disequilibrium Test (QPDT) to detect linkage disequilibrium between a genetic marker and quantitative traits in general pedigrees. For the 12th Genetic Analysis Workshop (GAW12), we apply the QPDT to detect association between genetic polymorphisms in the seven candidate genes and five quantitative traits in the simulated data sets. With reasonable power, we can identify genes that are associated with the quantitative traits. When the PDT is applied to identify linkage between genetic polymorphisms and the qualitative trait, we can only detect the candidate gene that directly affects the qualitative trait.

METHODS

The PDT was described in Martin et al. [2000], and we only describe the QPDT here. The QPDT utilizes information from the following three types of nuclear families in an extended pedigree: (I) Families with both parents available and at least one parent is heterozygous at the marker being studied; (II) Families with one available parent and one or more offspring where all the offspring have the same genotypes; and (III) Families with at most one available parent and multiple offspring where at least two siblings have different genotypes. An extended pedigree is informative if it contains at least one nuclear family with one of the above three structures.

For a nuclear family, let Y_i denote the trait value of the i th child for a quantitative trait of interest. Assume the genetic marker being studied is biallelic with two alleles: M and N. Let X_i denote the number of M alleles carried by the i th child and \bar{X} denote the mean number of M alleles among all the offspring in this nuclear family. For the first type of nuclear family that has both parents and at least one parent is heterozygous at the marker, define $X_{im} = 1$ (or -1) if the mother is heterozygous and transmits allele M (or N) to the i th child, and $X_{im} = 0$ if the mother is homozygous. We similarly define X_{if} for the father. For the second type of nuclear family that has one available parent and one or more offspring where all the offspring have the same genotypes, we only consider offspring-parent pairs with genotypes (NM, NN) or (MM, NM), and offspring-parent pair with genotypes (NN, NM) or (NM, MM). The first genotype in the bracket is the offspring's genotype and second genotype in the bracket is the available parent's genotype. Using the notation in Sun et al. [1999, 2000], we define $X_{(1)} = 1$ if the genotypes for the offspring-parent pair are (NM, NN) or (MM, NM), $X_{(1)} = -1$ if the genotypes for the parent-offspring pair are (NN, NM) or (NM, MM), and $X_{(1)} = 0$ for other genotypes of the offspring-parent pair. Define random variables U_1, U_2 and U_3 as

the covariance between the trait values and the genotypes for the first, second, and third types of nuclear family:

$$U_1 = \sum_{i=1}^t (Y_i - \bar{Y})(X_{im} + X_{if}), \quad U_2 = \sum_{i=1}^t (Y_i - \bar{Y})X_{(1)}$$

and

$$U_3 = \sum_{i=1}^t (Y_i - \bar{Y})(X_i - \bar{X}), \quad (1)$$

where t is the number of offspring in this nuclear family and \bar{Y} is the mean trait value of all of the offspring in all of the pedigrees. Under the null hypothesis of no linkage disequilibrium, $E(U_k) = 0$ for $k = 1, 2$, and 3.

For an extended pedigree, let n_1 , n_2 , and n_3 denote the number of the first, second, and third types of nuclear families, respectively. Define

$$D = \frac{1}{n_1 + n_2 + n_3} \left(\sum_{j_1=1}^{n_1} U_{j_1,1} + \sum_{j_2=1}^{n_2} U_{j_2,2} + \sum_{j_3=1}^{n_3} U_{j_3,3} \right), \quad (2)$$

where $U_{j_1,1}$, $U_{j_2,2}$, and $U_{j_3,3}$ are the covariances between the trait values and the genotypes for the j_k -th nuclear family of the k -th type. For n independent extended pedigrees, let D_l denote the random variable D defined for the l -th extended pedigree, then under the null hypothesis of no association,

$$\mu = E\left(\sum_{l=1}^n D_l\right) = 0, \text{ and}$$

$$\sigma^2 = \text{Var}\left(\sum_{l=1}^n D_l\right) = \sum_{l=1}^n \text{Var}(D_l) = E\left(\sum_{l=1}^n D_l^2\right).$$

Hence, if we define the test statistic as

$$T = \frac{\sum_{l=1}^n D_l}{\sqrt{\sum_{l=1}^n D_l^2}}, \quad (3)$$

Under the null hypothesis of no linkage disequilibrium, T is asymptotically normally distributed with mean 0 and variance 1. Therefore, we propose to use T in (3) as the QPDT statistic, and the statistical significance for association can be estimated by comparing the observed T to a standard normal distribution.

RESULTS

Association Tests

We apply the QPDT as well as the PDT statistics to analyze the sequence data from the seven candidate genes in the simulated data sets in GAW12. All of the 50 replications of the simulated data sets in the isolated population are used. For each replication, the data consist of 23 pedigrees with 1497 individuals. For each candidate gene, we apply the QPDT to detect association between each individual polymorphism and all five quantitative traits within the gene. Only polymorphic markers whose major allele

frequency is less than 95% are used. The number of the polymorphic markers used in gene 1 to gene 7 is approximately 100, 50, 55, 105, 30, 30, and 130, respectively. Note that the exact number of markers for each gene varies from replication to replication. In addition, we apply the PDT [Martin et al., 2000] to detect association between each individual polymorphism and the qualitative trait. We estimate the overall statistical significance level for a given gene by the smallest p-value achieved among all polymorphic markers with Bonferroni correction. We use the 50 replications to estimate the power of the QPDT for association between the genes and the quantitative traits and the power of the PDT for association between the genes and the qualitative trait. The results are summarized in Table I.

Table I: The power of the QPDT for the associations between the seven candidate genes and the five quantitative traits and the power of the PDT for the associations between the seven candidate genes and the qualitative trait at statistical significance level 5% after Bonferroni correction. Q_1, \dots, Q_5 represent the five quantitative traits.

Gene	Quantitative traits					Qualitative Trait
	Q_1	Q_2	Q_3	Q_4	Q_5	
1	0	0	0	0	0	0.92
2	0	0	0	0	0.16	0.02
3	0	0	0	0	0.02	0.02
4	0	0	0	0	0	0
5	0	0.02	0	0	0	0.02
6	0.64	0.26	0	0	0	0.03
7	0	0	0	0	0	0

Genes 1, 3, 4, 5, and 7 have no association with the five quantitative traits. The powers of the QPDT for the associations between these five genes and the five quantitative traits are almost all zero. Therefore, the Bonferroni correction results in conservative tests, likely due to the strong linkage disequilibrium among the markers being studied. Gene 2 is associated with quantitative trait 5 and has no association with all other four quantitative traits. The association between Gene 2 and quantitative trait 5 is detected in 16% of the replications and no association is detected between Gene 2 and other traits. The low power is possibly due to the fact that there are multiple functional alleles at different locations within Gene 2. Gene 6 is associated with quantitative traits 1 and 2. The association between Gene 6 and quantitative trait 1 is detected in 65% of the replications and the association between Gene 6 and quantitative trait 2 is detected in 26% of the replications. Again, no false positive associations are detected between Gene 6 and the other three traits.

For the qualitative trait, the PDT detects the association between the qualitative trait and Gene 1, which directly affects the qualitative trait, in 92% of the replications. However, the PDT has little power to detect the association between the qualitative trait and Gene 2 and Gene 6, which indirectly affect the qualitative trait.

Location Estimation

Within a given gene, the location of the genetic variant can be estimated to be at the polymorphic marker that yields the smallest p-value. In our analysis, we only estimate the location if the smallest p-value is less than 10% for a certain trait. In Figures 1(a) and 1(b), we plot the distribution of the estimated locations of genetic variants associated with

the quantitative traits in Gene 6 and Gene 2, respectively, on the basis of 50 replications. It can be seen that the estimated location of the trait locus in Gene 6, between sites 8000 and 8500, is $\sim 2.5\text{kb}$ from the functional mutation at site 5782. This bias may be caused by very strong linkage disequilibrium among the polymorphic markers within this gene. The estimated location of the trait locus in Gene 2 is more spread out among the 50 replications, due to the facts that there are multiple functional alleles throughout this gene. In Figures 1(c) and 1(d), we plot the distribution of the estimated locations of genetic variants associated with the qualitative trait in Gene 6 and Gene 1, respectively. The estimated genetic variant location in Gene 6 is very close to the true location for half of the replications and spread out for the rest of the replications. Compared to the distribution of the estimated locations in Gene 6, there is a much higher variation in the estimated locations in Gene 1. For the other gene-trait combinations not shown in Figure 1, the estimated p-value is less than 10% in only one or two replications among the 50 replications.

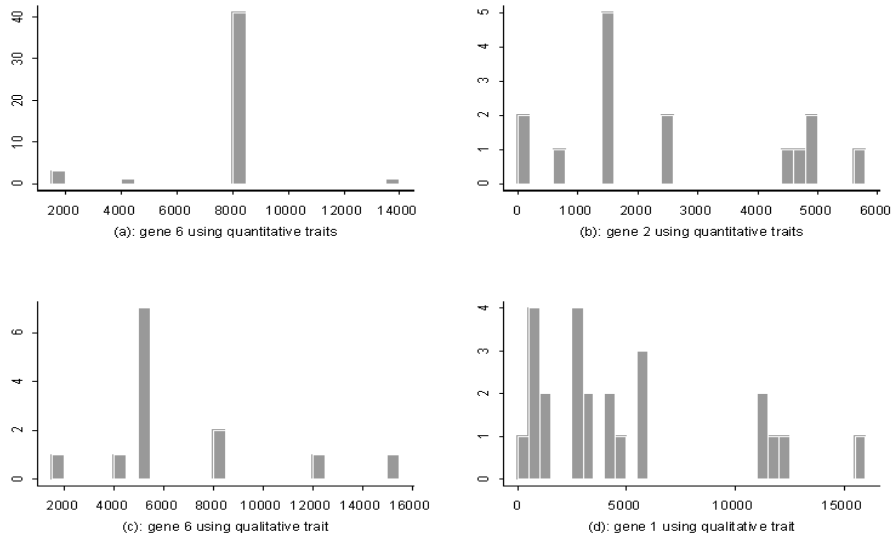


Fig. 1. The distributions of the estimated genetic variant locations among the 50 replications in genes associated with the trait of interest.

DISCUSSION

The motivation for developing the QPDT was to analyze quantitative traits using the large pedigree data in GAW12. Although there are many tests available for independent nuclear families, there has been little discussion of association testing in large pedigrees. Standard tests require selection of a single nuclear family or discordant sibship from extended pedigrees. Clearly, this is less than optimal, since it discards data. Furthermore, results using a set of extended pedigrees may vary because of random selection of nuclear families or sibships for inclusion. Martin et al. [2000] proposed the PDT to detect linkage disequilibrium between a genetic marker and qualitative traits in general pedigrees. The PDT uses all the nuclear families with both parents available and nuclear families with multiple offspring of different genotypes. The QPDT proposed in this paper to detect

linkage disequilibrium between a genetic marker and quantitative traits in general pedigrees is a generalization of the PDT to quantitative traits. Comparing with the PDT, besides the two kinds of nuclear families used by the PDT, the QPDT uses the nuclear families with one parent and one offspring or multiple offspring with same genotype as well. Comparing with the methods proposed by Sun et al. [1999, 2000], which are only applicable to independent nuclear families, the QPDT can use related nuclear families extracted from extended pedigrees in the same analysis.

When we applied the QPDT to analyze the sequence data from the seven candidate genes in the simulated data sets in GAW12, we can identify genes that are associated with the quantitative traits with reasonable power. When the PDT is applied to identify linkage disequilibrium between genetic polymorphisms and the qualitative trait, we can only detect the candidate gene that directly affects the qualitative.

Since there are many polymorphic markers within a gene in the sequence data and we test the association between trait values and marker genotypes marker by marker, we use the Bonferroni adjustment to evaluate the p-value to counter the multiple comparison problem. However, the Bonferroni method is very conservative. We can see from our results (Table I) that the power of the QPDT for the associations between the genes that have no association with quantitative traits and the five quantitative traits is almost all zero at the 5% significance level. An alternative method to control the false discovery rate (FDR) [Benjamini and Hochberg, 1995] may be more appropriate in many cases. The FDR is defined to be the number of false rejections divided by the number of total rejections. Suppose there are a total of m tests in a multiple testing problem. Let $p_{(1)}, \dots, p_{(m)}$ be the ordered p-values. For $i^* = \max\{i, p_{(i)} \leq (i/m)\alpha\}$, we reject all the hypotheses whose p-values are $p_{(1)}, \dots, p_{(i^*)}$, respectively. Let R be the number of hypotheses rejected and R^* the number of false rejections so that $Q = R^*/R$ is the false discovery rate. If the above procedure is followed, then Benjamini and Hochberg [1995] proved that $E(Q) \leq \alpha$. The usefulness of this approach in complex disease gene mapping needs to be further investigated.

ACKNOWLEDGEMENTS

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

REFERENCES

- Abecasis GR, Cardon LR, Cookson WO (2000): A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279-292.
- Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 57:289-300.
- Martin ER, Monks SA, Warren LL, Kaplan NL (2000): A test for linkage and association in General Pedigree: The Pedigree Disequilibrium Test. *Am J Hum Genet* 67:146-154.
- 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 Q, Khoury MJ (1999): Transmission disequilibrium test (TDT) when only one parent is available: The 1-TDT. *Am J Epidemiol* 150:97-104.
- Sun FZ, Flanders WD, Yang Q, Zhao H (2000): Transmission/disequilibrium test for quantitative traits. *Ann Hum Genet*, in press.
- Zhao H (2000): Family-based association designs. *Stat Method Med Res*, in press.