RESEARCH ARTICLE

Statistical equivalent of the classical TDT for quantitative traits and multivariate phenotypes

TANUSHREE HALDAR and SAURABH GHOSH*

Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India

Abstract

Clinical end-point traits are usually governed by quantitative precursors. Hence, there is active research interest in developing statistical methods for association mapping of quantitative traits. Unlike population-based tests for association, family-based tests for transmission disequilibrium are protected against population stratification. In this study, we propose a logistic regression model to test the association for quantitative traits based on a trio design. We show that the method can be viewed as a direct extension of the classical transmission diequilibrium test for binary traits to quantitative traits. We evaluate the performance of our method using extensive simulations and compare it with an existing method, family-based association test. We found that the two methods yield comparable powers if all families are considered. However, unlike FBAT, which yields an inflated rate of false positives when noninformative trios with all three individuals' heterozygous are removed, our method maintains the correct size without compromising too much on power. We show that our method can be easily modified to incorporate multivariate phenotypes. Here, we applied this method to analyse a quantitative endophenotype associated with alcoholism.

[Haldar T. and Ghosh S. 2015 Statistical equivalent of the classical TDT for quantitative traits and multivariate phenotypes. *J. Genet.* **94**, 619–628]

Introduction

The classical transmission disequilibrium test (TDT) for binary traits proposed by Spielman *et al.* (1993) is a family-based alternative to population-based case–control studies and circumvents the problem of population stratification as it tests for allelic association in the presence of linkage. However, most binary traits relate to clinical end-points that are defined by thresholds of quantitative precursors. Since quantitative traits (QT) carry more information within genotype variations, it has been argued that it may be a more prudent strategy to analyse the quantitative phenotypes without dichotomizing them into binary traits. Hence, it is current research interest to explore if the classical TDT can be modified for association analyses of QT. The paradigm of family-based association in the context of QT is not very straight-forward and methods (Allison 1997; Rabinowetz 1997; George *et al.* 1999; Abecasis *et al.* 2000; Monks and Kaplan 2000; Lange and Laird 2002) have generally considered the intuitive concept of differences in allelic transmissions between offspring having high QT values and those having low values as evidence of linkage disequilibrium (LD) for QT. However, many of these tests suffer from some inherent limitations: they are often sensitive to violations in distributional assumptions of the QT such as normality, assume a linear relationship between allele transmission and the QT and/or use selected sampling resulting in loss of information and power. A detailed review of the different tests is available in Ewens *et al.* (2008). Moreover, these methods are not direct extensions of the classical TDT. We propose a computationally simple logistic regression-based test that can be analytically shown to be statistically equivalent to the TDT for binary traits, and hence is not susceptible to the presence of population stratification in the data. We perform Monte-Carlo simulations under a wide spectrum of genetic models and probability distributions of the QT to assess the power of the proposed procedure. We compare the performance of our proposed method with a similar model-free method, FBAT (Lange and Leird 2002), which is computationally more intensive. We evaluate the rates of false positives and powers of both the methods when trios with all three individuals' heterozygous are removed from the analyses and only one transmission is considered

*For correspondence. E-mail: saurabh@isical.ac.in.

Keywords. family-based genetic association; logistic regression; linkage disequilibrium.

for trios with both parents heterozygous and the offspring homozygous. The proposed method can be easily extended to incorporate multivariate phenotypes. We apply our method to analyse externalizing symptoms, an alcoholism-related endophenotype from the Collaborative Study on the Genetics Of Alcoholism (COGA) project.

Model and data description

We assume that a QT Y is controlled by a biallelic QT locus (QTL) with alleles A and a. We consider a biallelic marker locus with alleles M_1 and M_2 such that the recombination fraction between the QTL and the marker locus is θ and the coefficient of LD between the loci is measured by $\delta = P(AM_1) - P(A)P(M_1)$, suppose that the allele frequencies of A and M_1 are p and m, respectively. We assume that the probability density of Y conditioned on the genotype at the QTL is f_1, if the genotype is AA, f_2, if it is Aa and f_3, if it is aa.

The data requirement is similar to the classical TDT for binary traits (Spielman *et al.* 1993). The data comprise marker genotypes of trios (two parents and an offspring) such that at least one of the parents is heterozygous and QT values of the offspring.

Statistical method

We propose a logistic link function to model the conditional distribution of a marker allele transmitted by a parent to an offspring given that the parent is heterozygous at the marker locus and the QT value y of the offspring.

$$P(Z = 1|y) = \exp(\beta(y - c))/\{1 + \exp(\beta(y - c))\},$$

$$P(Z = 0|y) = 1/\{1 + \exp(\beta(y - c))\},$$

where c is some central summary measure of the QT Y in the population such as mean or median,

$$Z = \begin{cases} 1 \text{ if } M_1 \text{ is transmitted from a heterozygous parent to an offspring,} \\ 0 \text{ if } M_2 \text{ is transmitted from a heterozygous parent to an offspring.} \end{cases}$$

We shall show that testing for H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$ in the above logistic model is equivalent to testing H_0: $\theta = 0.5$ (no linkage) or $\delta = 0$ (no association) vs H_1: $\theta < 0.5$ and $\delta \neq 0$ (both linkage and association) and hence, equivalent to the classical TDT (Spielman *et al.* 1993) when extended to QT.

Proof of the equivalence

The joint probabilities of transmission and nontransmission of marker alleles M_1 and M_2 by a parent given that the different genotypes of the offspring at the QTL are provided in appendix.

We first note that f_1, f_2 and f_3 being the probability densities of Y conditioned on the QTL genotypes, it is not possible that $f_1(y) = f_2(y) = f_3(y)$ for all y.

$$\begin{aligned}
& P(Z = 1|y) \\
&= P \text{ (transmitting } M_1 \text{ by a heterozygous parent } |y), \\
&= P \text{ (transmitting } M_1 \text{ and not transmitting } M_2 \text{ by a parent } |y), \\
&= \sum\nolimits_G \text{P (transmitting } M_1 \text{ and not transmitting } M_2 \text{ by a parent, genotype of the offspring is } G|y), \text{where the sum is over all possible QTL genotypes } G = AA, Aa \text{ and } aa. \\
&= \sum\nolimits_G P(Z = 1|G)P(G|y), \\
&= \sum\nolimits_G P(Z = 1|G)P(G)f(y|G)/\left\{\sum\nolimits_G P(G)f(y|G)\right\}, \\
&= m(1-m)+(1-\theta-m)\delta\{pf_1(y)+(1-2p)f_2(y)-qf_3(y)\}/ \\
&\quad \{p^2f_1(y) + 2pqf_2(y) + q^2f_3(y)\}.
\end{aligned}$$

Using similar conditioning arguments:

$$\begin{aligned}
P(Z = 0|y) = m(1-m) + (\theta - m)\delta\{pf_1(y) + (1 - 2p)\, f2(y) \\
-qf_3(y)\}/\{p^2f_1(y) + 2pqf_2(y) + q^2f_3(y)\}
\end{aligned}$$

where $q = 1-p$. We note that only the $(1, 2)$th and the $(2, 1)$th elements of each of the probability tables in appendix are required in calculating these probabilities.

In the logistic model:

$$\begin{aligned}
& \beta = 0 \\
\Leftrightarrow\ & P(Z = 1|y) = P(Z = 0|y) = 0.5, \\
\Leftrightarrow\ & P(Z = 1|y) - P(Z = 0|y) = 0, \\
\Leftrightarrow\ & [m(1 - m) + (1 - \theta - m)\delta\{pf_1(y) + (1 - 2p)f_2(y) \\
& -qf_3(y)\}/\{p^2f_1(y) + 2pqf_2(y) + q^2f_3(y)\}] - [m(1 - m) \\
& +(\theta - m)\delta\{pf_1(y) + (1 - 2p)f_2(y) - qf_3(y)\}/ \\
& \{p^2f_1(y) + 2pqf_2(y) + q^2f_3(y)\}] = 0, \\
\Leftrightarrow\ & (1 - 2\theta)\delta\{pf_1(y) + (1 - 2p)f_2(y) - qf_3(y)\}/\{p^2f_1(y) \\
& +2pqf_2(y) + q^2f_3(y)\} = 0, \\
\Leftrightarrow\ & (1 - 2\theta)\delta\{pf_1(y) + (1 - 2p)f_2(y) - qf_3(y)\} = 0, \\
\Leftrightarrow\ & (1 - 2\theta)\delta = 0, \text{ since the condition has to hold for all } y \\
& \text{and there exists } y \text{ such that } f_1(y),\ f_2(y) \text{ and } f_3(y) \text{ are not equal.} \\
\Leftrightarrow\ & \theta = 0.5 \text{ or } \delta = 0 \text{ (identical to the null hypothesis of the classical TDT).}
\end{aligned}$$

Test procedure

Based on the equivalence, the test for no linkage or no association can be performed by a likelihood-ratio test on β. We use the sample mean or median of Y, as an estimator of c in the proposed logistic model. Suppose there are n transmissions from heterozygous parents, the log-likelihood statistic can be expressed as $2\log_e L(\hat{\beta}) + n\log_e 4$, where $\hat{\beta}$ is the unrestricted maximum likelihood estimator (m.l.e.) of β and

is asymptotically distributed as chi-squares with one degree of freedom under the null hypothesis.

The major advantage of family-based designs over population-based studies for detecting allelic association is the property of family-based tests being protected against inflated rates of false positives generated by possible population substructure. We show that the proposed test is not susceptible to population stratification with respect to false positives. Suppose there are k subpopulations with varying allele frequencies at a marker locus which is unlinked to the QTL in each of the subpopulations. We first observe that the test procedure does not involve estimation of allele frequencies. If the log-likelihood function for the ith subpopulation is $l_i(\beta), i = 1, 2, \ldots, k$, then the overall log-likelihood function is given by $l(\beta) = \sum_{i=1}^{k} l_i(\beta)$. Since the marker locus is unlinked to the QTL in each of the subpopulations, $l_i(\beta)$ is maximized at $\beta = 0, \forall i = 1, 2, \ldots, k$. Hence, $\sum_{i=1}^{k} l_i(\beta)$ is also maximized at $\beta = 0$ implying that the proposed test does not yield false positives due to heterogeneity in allele frequencies across subpopulations.

Here we note that there are competing choices of modelling allelic transmissions in trios comprising two heterozygous parents. For example, a heterozygous offspring (with genotype M_1M_2) would imply that one parent has transmitted the M_1 allele while the other parent has transmitted the M_2 allele. The inclusion of such a trio may result in reduced power as it is not informative about linkage and association. On the other hand, ignoring such a trio may result in inflated false positive rates. Thus, there is a choice between ignoring such a trio and including it. Similarly, if the offspring is homozygous, it would mean that both the parents have transmitted the same allele to the offspring and would result in the same likelihood in the proposed logistic model for either of the transmissions. Thus, there is a choice between considering only one transmission versus two identical transmissions for such a trio. We have evaluated the relative performances of all the competing choices.

Simulations and results

To assess the power of the proposed test, we perform simulations under different genetic models and probability distributions of the underlying QT. Details of the simulation steps are provided in appendix. Data were generated on 500 trios in each set of simulations. We chose the genetic parameters such as the QTL and marker allele frequencies, the QT means and variances conditioned on the QTL genotypes such that the proportion of trait variance explained by the QTL ranged between 5 and 20%. We considered three different probability distributions of the QT: normal (reflecting symmetric QT distributions), location-shifted chi-squares (reflecting skewness in QT distributions) and location-shifted Poisson (reflecting discrete phenotypes such as symptom counts associated with a clinical manifestation). We refer our proposed method as transmission-based association test (TBAT) and compare its performance with FBAT (Lange and Laird 2002), which is based on a standardized covariance measure between the quantitative phenotype and the count of one of the alleles in an offspring. The primary difference between TBAT and FBAT pertains to the modelling of trios with both parents heterozygous. For such families, FBAT models the joint probability of the allelic transmissions from the two parents conditioned on their genotypes, but TBAT considers the allelic transmission from each parent independently.

We find that the results are similar for the different probability distributions of the QT. This is expected as the proposed logistic model is conditioned on QT values and hence, robust to the underlying distribution of the QT. We present the results in table 1, when the QT has a normal distribution and there is no dominance at the QTL. For brevity, the results for the other two distributions, being similar to table 1, are not presented separately. The choice of the summary central measure as the sample mean or the median does not have differential impact on the results. When all transmissions from heterozygous parents are considered (i.e., our method considers two transmissions for trios with both parents heterozygous), we find that our proposed method as well as FBAT maintain the correct size. When trios with both parents heterozygous are removed from the analyses, both the methods have inflated false positive rates. A similar phenomenon is observed when trios with all three individuals' heterozygous are ignored, but two transmissions are considered for trios with both parents heterozygous and the offspring homozygous. However, if we consider only one transmission for such trios (henceforth referred as the 'restricted set' of trios), we find that while our method has the correct size, FBAT has inflated false positive rates. The false positive rate increases with heterozygosity at the marker locus. Thus, the powers of the two methods can be compared only when the analyses are based on all transmissions from heterozygous parents. We find that the empirical powers of our proposed method are comparable to those based on FBAT irrespective of the distribution of the QT. The power of our method increases with increase in dominance (γ) at the QTL. This can be explained by the fact that the proportion of trait variation explained by the QTL is an increasing function of the dominance. However, models with different levels of dominance at the QTL, but explaining the same proportion of trait variation yields similar powers. We also find that for a given allele frequency at the QTL, the power is maximum when the marker allele frequency is similar to that at the QTL. Thus, it follows that SNPs that have less heterozygosity may be more useful in identifying rare variants. Compared to analyses based on all available transmissions from heterozygous parents, we find that the analyses based on the 'restricted set' of trios suffer only a marginal loss of power (less than 0.04). The difference in the powers become minimal in the presence of a rare variant at the QTL.

To evaluate the effect of dichotomizing the QT into a binary trait on the power of the proposed method, we used sample mean and sample median as two possible thresholds.

Table 1. Relative comparisons of empirical powers of the proposed test (TBAT) and FBAT for a normally distributed QT.

p	V (%)	δ*	m=0.1		m=0.3		m=0.5	
			TBAT	FBAT	TBAT	FBAT	TBAT	FBAT
All trios								
0.3	19.9	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.21	0.21	0.62	0.62	0.32	0.32
		0.66	0.64	0.64	1.00	1.00	0.85	0.85
		1	0.95	0.95	1.00	1.00	1.00	1.00
0.3	10.2	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.13	0.13	0.37	0.37	0.19	0.19
		0.66	0.38	0.37	0.91	0.91	0.57	0.57
		1	0.72	0.72	1.00	1.00	0.90	0.90
0.05	5.0	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.12	0.12	0.07	0.07	0.06	0.06
		0.66	0.34	0.33	0.13	0.13	0.08	0.08
		1	0.64	0.64	0.23	0.23	0.12	0.12
Restricted trios								
0.3	19.9	0	0.05	0.06	0.05	0.08	0.05	0.09
		0.33	0.21		0.59		0.30	
		0.66	0.63		0.99		0.81	
		1	0.94		1.00		0.99	
0.3	10.2	0	0.05	0.06	0.05	0.08	0.05	0.09
		0.33	0.13		0.34		0.17	
		0.66	0.37		0.88		0.53	
		1	0.70		1.00		0.87	
0.05	5.0	0	0.05	0.06	0.05	0.08	0.05	0.09
		0.33	0.12		0.07		0.06	
		0.66	0.33		0.12		0.08	
		1	0.63		0.21		0.12	

p, Minor allele frequency at QTL; m, minor allele frequency at marker locus; δ*, LD parameter; V, percentage of variance explained by QTL.

We found that for both choices, the powers are significantly lower compared to those obtained for the QT. This is expected as the information within genotype variability is reduced if the QT values are dichotomized. While the two thresholds yield comparable powers when the QT values have a normal distribution, the sample median performs marginally better than the mean when the distribution of the QT is skewed. We also observe that the powers obtained using our proposed method and FBAT are similar even when the QT is dichotomized.

Selected sampling

Owing to phenotyping constraints, data on quantitative precursors are often collected only on individuals with a clinical manifestation (for e.g., anti-CCP or IgM in studies on rheumatoid arthritis) and hence, such data may not reflect the true probability distribution of the QT. Moreover, there is both analytical and empirical evidences that sampling from the two tails of the underlying distribution of a QT may result in increased power in detecting association in population-based studies (Abecasis *et al.* 2001; Chen and Li 2011). This is intuitively expected as the allele frequencies at the QTL will differ significantly in the two extremes of the QT distribution. Thus, it is of interest to assess the effect of selected sampling on powers of family-based designs such as our proposed method and FBAT. Assuming that the QT has a a normal distribution, we carried out simulations using the same parameters as earlier, under two sampling schemes: (i) selecting offspring with QT values higher than the 0.8 quantile (equivalent to selecting affected offspring in a disease model with prevalence 0.2) and (ii) selecting individuals with QT values lower than the 0.1 quantile or higher than the 0.9 quantile (i.e., 10% from each tail of the QT distribution). Since only 20% of a random set of informative trios are selected for the analyses, an initial sample of 500 trios may result in an insufficient number of observations to ensure the asymptotic chi-squares distribution of the test statistic under the null and thus, yield an inflated false positive rate. We varied the number of trios from 500 to 1000 in increments of 100 in our simulations and observed that while 500 trios were sufficient to maintain the correct size of the test, when the heterozygosity at the marker locus is high ($m = 0.3$ and 0.5) and the analyses are based on all informative trios, a sample size of 800 was required when the 'restricted set' of trios was considered. When the heterozygosity is low ($m = 0.1$), 1000 trios were necessary for the test to have the correct size. The results based on a sample size of 1000 trios are provided in table 2. When the selected sampling is only at the upper tail, the distribution of the QT values is extremely skewed and hence, we find that the choice of sample median as the summary

Table 2. Relative comparisons of empirical powers of the proposed test (TBAT) and FBAT for selected sampling on a normally distributed QT.

	m=0.1				m=0.3				m=0.5			
	TBAT		FBAT		TBAT		FBAT		TBAT		FBAT	
δ*	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Only upper 20%, all trios												
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.33	0.06	0.08	0.06	0.07	0.08	0.15	0.08	0.14	0.07	0.10	0.07	0.10
0.66	0.07	0.13	0.07	0.12	0.17	0.41	0.17	0.41	0.12	0.24	0.11	0.24
1	0.09	0.21	0.09	0.20	0.31	0.73	0.30	0.72	0.21	0.48	0.20	0.48
Only upper 20%, restricted trios												
0	0.05	0.05	0.06	0.06	0.05	0.05	0.08	0.08	0.05	0.05	0.09	0.09
0.33	0.06	0.07			0.08	0.14			0.07	0.09		
0.66	0.07	0.13			0.14	0.37			0.11	0.22		
1	0.08	0.20			0.23	0.66			0.18	0.43		
Upper 10% and lower 10%, all trios												
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
0.33	0.15	0.11	0.15	0.11	0.45	0.31	0.45	0.31	0.23	0.17	0.23	0.17
0.66	0.46	0.32	0.45	0.31	0.96	0.83	0.96	0.83	0.68	0.51	0.68	0.50
1	0.81	0.66	0.81	0.65	1.00	1.00	1.00	0.99	0.96	0.85	0.96	0.85
Upper 10% and lower 10%, restricted trios												
0	0.05	0.05	0.06	0.06	0.05	0.05	0.08	0.08	0.05	0.05	0.09	0.09
0.33	0.15	0.11			0.42	0.29			0.21	0.15		
0.66	0.44	0.31			0.94	0.79			0.63	0.46		
1	0.80	0.63			1.00	0.99			0.94	0.81		

m, Minor allele frequency at marker locus; δ*, LD parameter; avg., average; med., median.

central measure yields more than twice the power compared to the sample mean. However, there is significant loss of power compared to sampling from the full distribution of the QT. On the other hand, when the selected sampling is performed at both tails of the QT distribution, we find that the sample mean outperforms the sample median with respect to power. Moreover, consistent with our expectation, there is a gain in power compared to sampling from the full distribution of the QT, even though the analyses are based on only 40% of the original sample size (20% of the number of informative trios among 1000 trios *vis-a-vis* the number of informative trios among 500 trios).

Extension to multivariate phenotypes

A single QT is often not a sufficiently good surrogate for a clinical end-point and it may be more optimal to consider a multivariate phenotype possibly comprising both quantitative and qualitative precursors of the end-point trait. For example, cholesterol levels, lipoprotein levels and systolic blood pressure may comprise a multivariate phenotype vector for cardiovascular disorder. Suppose data are available on a vector of k traits $Y_1, Y_2, \ldots, Y_k$. Given data on genotypes at a biallelic marker for a trio with at least one parent heterozygous and the multivariate phenotype of the offspring, we consider a logistic regression model in similar lines as the univariate set-up described earlier:

$$P(Z=0|y_1,y_2,\ldots,y_k) = 1/\exp\left(\sum_{i=1}^{k}\beta_i(y_i-c_i)\right),$$

$$P(Z=1|y_1,y_2,\ldots,y_k) = \exp\left\{\sum_{i=1}^{k}\beta_i(y_i-c_i)\right\} \Big/ \left\{1+\exp\left(\sum_{i=1}^{k}\beta_i(y_i-c_i)\right)\right\},$$

where c_i is some central summary measure of the QT Y_i in the population,

$$Z = \begin{cases} 1 & \text{if } M_1 \text{ is transmitted from a heterozygous parent to an offspring,} \\ 0 & \text{if } M_2 \text{ is transmitted from a heterozygous parent to an offspring.} \end{cases}$$

The test for no linkage or no association versus the presence of both linkage and association is equivalent to testing H_0: $\beta_i = 0 \forall i = 1,2,\ldots,k$ vs. H_1: $\beta_i \neq 0$ for at least one i. Using the sample mean or median of y_i as an estimator of c_i in the logistic model, the standard log-likelihood ratio test statistic is distributed as chi-squares with k degrees of freedom under the null hypothesis.

Using simulations, we compare the empirical power of our method with FBAT for two bivariate phenotype models: (i) two QT distributed as bivariate normal and (ii) one QT and one binary trait defined by dichotomizing a QT. We evaluated the powers of the proposed test based on logistic regression for different levels of correlation between the two constituent traits in the bivariate phenotype vector. The different choices of the simulation parameters are provided in appendix. The

results pertaining to the first model are provided in table 3 and those pertaining to the second model are provided in table 4 for two values of the correlation coefficient between the two traits: 0 and 0.3. For both models, we observe that the power of the test decreases with increase in the extent of correlation between the two traits. This can be explained by the fact that lower the correlation between the traits, more would be the information carried by each trait independently. We also find that genetic models with the same proportion of trait variation explained by the QTL and the same correlation coefficient between the two traits yield almost identical powers (the results are not presented for brevity). When we compare the power of our proposed method with FBAT, we obtain similar inferences as those discussed for the univariate model.

Application to COGA phenotype

Collaborative study on the genetics of alcoholism (COGA) is a multicentre research programme established to detect and map susceptibility genes for alcohol-dependence and related phenotypes. A well-known link exists between externalizing behaviour disorders such as conduct disorder and alcohol-use disorders. Ghosh *et al.* (2008) performed a genomewide linkage scan on a quantitative endophenotype defined as

Table 3. Relative comparisons of empirical powers of the proposed test (TBAT) and FBAT for a bivariate phenotype comprising two normally distributed QT.

			m=0.1		m=0.3		m=0.5	
p	V (%)	δ*	TBAT	FBAT	TBAT	FBAT	TBAT	FBAT
				$P = 0.0$				
All trios								
0.3	10.20	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.17	0.16	0.53	0.53	0.25	0.25
		0.66	0.54	0.54	0.99	0.99	0.78	0.78
		1	0.91	0.91	1.00	1.00	0.99	0.99
0.05	5.00	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.15	0.14	0.08	0.07	0.06	0.06
		0.66	0.47	0.46	0.16	0.15	0.09	0.09
		1	0.83	0.82	0.31	0.31	0.16	0.16
Restricted trios								
0.3	10.20	0	0.05	0.06	0.05	0.09	0.05	0.11
		0.33	0.16		0.45		0.23	
		0.66	0.53		0.99		0.74	
		1	0.90		1		0.98	
0.05	5.00	0	0.05	0.07	0.05	0.09	0.05	0.11
		0.33	0.15		0.07		0.06	
		0.66	0.46		0.15		0.09	
		1	0.82		0.29		0.15	
				$P = 0.3$				
All trios								
0.3	10.20	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.14	0.13	0.42	0.42	0.20	0.20
		0.66	0.44	0.43	0.96	0.96	0.66	0.66
		1	0.82	0.81	1.00	1.00	0.96	0.96
0.05	5.00	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.12	0.12	0.07	0.07	0.06	0.06
		0.66	0.38	0.37	0.13	0.13	0.09	0.08
		1	0.72	0.72	0.25	0.24	0.13	0.13
Restricted trios								
0.3	10.20	0	0.05	0.06	0.05	0.09	0.05	0.11
		0.33	0.14		0.39		0.19	
		0.66	0.42		0.95		0.61	
		1	0.80		1		0.94	
0.05	5.00	0	0.05	0.07	0.05	0.09	0.05	0.11
		0.33	0.12		0.07		0.06	
		0.66	0.37		0.13		0.08	
		1	0.71		0.23		0.12	

p, Minor allele frequency at QTL; m, minor allele frequency at marker locus; δ*, LD parameter; V, percentage of variance explained by QTL; *P*, correlation between two traits.

Table 4. Relative comparisons of empirical powers of the proposed test (TBAT) and FBAT for a bivariate phenotype comprising a normally distributed QT and a binary trait.

p	V (%)	δ*	m=0.1		m=0.3		m=0.5	
			TBAT	FBAT	TBAT	FBAT	TBAT	FBAT
				$P = 0.0$				
All trios								
0.3	10.20	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.13	0.12	0.39	0.39	0.19	0.19
		0.66	0.40	0.39	0.94	0.94	0.63	0.62
		1	0.76	0.75	1.00	1.00	0.95	0.95
0.05	5.00	0	0.06	0.05	0.05	0.05	0.05	0.05
		0.33	0.13	0.11	0.07	0.07	0.06	0.06
		0.66	0.36	0.34	0.13	0.13	0.09	0.08
		1	0.69	0.68	0.25	0.24	0.13	0.13
Restricted trios								
0.3	10.20	0	0.05	0.06	0.05	0.09	0.05	0.11
		0.33	0.13		0.36		0.18	
		0.66	0.38		0.92		0.58	
		1	0.74		1		0.92	
0.05	5.0	0	0.06	0.06	0.05	0.09	0.05	0.10
		0.33	0.13		0.07		0.06	
		0.66	0.36		0.13		0.08	
		1	0.68		0.23		0.13	
				$P = 0.3$				
All trios								
0.3	10.20	0	0.05	0.05	0.05	0.05	0.05	0.05
		0.33	0.12	0.11	0.32	0.32	0.16	0.16
		0.66	0.33	0.32	0.88	0.88	0.52	0.52
		1	0.66	0.65	1.00	1.00	0.89	0.89
0.05	5.00	0	0.06	0.05	0.05	0.05	0.05	0.05
		0.33	0.12	0.10	0.07	0.06	0.06	0.06
		0.66	0.30	0.28	0.12	0.11	0.08	0.07
		1	0.58	0.56	0.20	0.20	0.11	0.11
Restricted trios								
0.3	10.20	0	0.05	0.06	0.05	0.09	0.05	0.11
		0.33	0.11		0.30		0.15	
		0.66	0.32		0.85		0.48	
		1	0.64		1		0.85	
0.05	5.00	0	0.06	0.06	0.05	0.09	0.05	0.10
		0.33	0.12		0.07		0.06	
		0.66	0.30		0.12		0.08	
		1	0.57		0.19		0.11	

p, Minor allele frequency at QTL; m, minor allele frequency at marker locus; δ*, LD parameter; V, percentage of variance explained by QTL; *P*, correlation between two traits.

the number of externalizing symptoms related to antisocial behavioural traits. The phenotype is the count of 24 symptoms endorsed by an individual and hence, ranges between 0 and 24. A multipoint nonparametric analysis based on 171 independent sibpairs revealed significant evidence of linkage in the 4q22.3 region on chromosome 4 harbouring the alcohol dehydorogenase (*ADH*) gene cluster. One of the linkage markers genotyped in this region was a biallelic marker, *ADH1C*. We performed the proposed logistic regression on 138 independent trios with at least one parent heterozygous at the locus. We found that one of the alleles is tramsmitted significantly more often ($P < 0.00001$) than the other allele for high values of the quantitative endophenotype, indicating that the marker locus may be in strong LD with a QTL modulating the number of externalizing symptoms.

Discussion

While population-based association studies are common in practice due to ease of data collection and statistical analyses, they are susceptible to population stratification. An alternative approach that circumvents this problem is the family-based design, the most popular being the classical TDT design (Spielman *et al.* 1993). In this study, we have developed a computationally simple test based on a logistic regression model for detecting association in the presence of

linkage. The classical TDT for binary traits can be viewed as a special case of the proposed method by coding the QT as a binary variable.

We have compared the performance of our method (TBAT) with a similar model-free method FBAT, that computes the likelihood of all transmissions within a trio simultaneously, while our method considers the likelihood of each transmission separately. Thus, the major difference in the two methods pertains to trios with both parents heterozygous at the marker locus. We find that both the methods yield comparable powers when all trios are considered. However, trios with all three individuals heterozygous do not contribute to increasing the power to detect linkage and association. When the analyses are based on the 'restricted set' of trios, we find that FBAT yields an inflated rate of false positives while our method maintains the correct size. Moreover, there is minimal loss of power in our method for the 'restricted set'. We also observe that the proportion of genetic variance explained by the QTL is the primary factor in determining the power of the proposed test, irrespective of the allele frequencies at the QTL and the marker locus as well as the gene effect sizes.

Since the proposed method is robust with respect to violations in distributional assumptions on the QT, our focus has been to compare it with other model-free alternatives. However, it is important to evaluate the power of our method relative to parametric alternatives, when relevant model assumptions are valid. Allison (1997) modelled the QT of an offspring as a function of his/her genotype conditioned on the parental mating type and used standard ANOVA to test the association under the assumption that the trait values are normally distributed. On similar lines, Abecasis *et al.* (2000) proposed QTDT that models the trait values based on linear effects of the offspring genotype and the average of the parental genotypes. We compared the performance of TBAT with that of QTDT when the underlying probability distribution of the QT values is (i) normal and (ii) chi-squares. Interestingly, we find that TBAT suffers a minimal loss in power compared to QTDT when the trait values are normally distributed. While, we have not separately presented the results based on QTDT due to brevity, we highlight some of the comparative powers of TBAT and QTDT. For example, the empirical power yielded by QTDT is 0.65 (the corresponding figure for TBAT is 0.62) when the minor allele frequencies at QTL and the marker locus are both 0.3, the proportion of variation in the trait explained by the QTL is 19.9% and the coefficient of LD is 0.33. Similarly, the empirical power is 0.39 (the corresponding figure for TBAT is 0.37) when the proportion of variation in the trait explained by the QTL is 10.2% and the other parameters are same as mentioned above. On the other hand, unlike TBAT which maintains the correct size of the test, QTDT produces inflated rates of false positives as high as 0.08 when the trait values are distributed as chi-squares. Moreover, QTDT becomes incompatible when analysing a multivariate phenotype comprising both quantitative as well as binary traits.

The alternative hypothesis of the proposed test is the presence of both linkage and association. Thus, the test is two-sided with respect to the choice of the marker allele. However, in the presence of some biological prior on the direction of the effect of an allele on QT values, one can perform a one-sided test for β using the logistic model. Since this results in a restricted alternative, the log-likelihood ratio test statistic will be asymptotically a 50 : 50 mixture of a degenerate distribution at 0 and a chi-squares distribution with one degree of freedom under the null hypothesis. Such a test is evidently more powerful and informative in detecting association compared to the proposed two-sided test.

The proposed method can be easily modified to analyse X-linked QT. However, since there cannot be any bias in the transmission of a paternal marker allele at a X-linked locus to an offspring, it is necessary that the mother in each trio is heterozygous at the marker locus. Thus, the same log-likelihood test procedure based on the logistic model can be carried out to test the association with X-linked QT.

Suppose, instead of the trio design, data are available on larger sibships. The proposed method can incorporate sibships by considering the transmissions from a heterozygous parent to the different sibs within a sibship to be independent. While such data are expected to carry more information on association in the presence of linkage, it has been argued that differential transmission of parental alleles at a marker locus among sibs with different QT values need not necessarily imply the presence of association as the marginal effect of linkage could result in the bias within a sibship.

We would like to emphasize that while family-based transmission disequilibrium tests are protected against inflated rates of false positives in the presence of population stratification, it is possible that they may be susceptible to an increase in the false negative rate. When genetically heterogeneous subpopulations are pooled together, it is possible that associations present in individual subpopulations may become insignificant. We are currently carrying out extensive simulations to evaluate the extent of adverse effect of population stratification on the false negative rates of family-based association tests.

A software TBAT incorporating the proposed test based on logistic regression is available on request.

Appendix

Joint probabilities of transmission and nontransmission of alleles M_1 and M_2 conditioned on QTL genotype AA

Trans. vs nontrans.	M_1	M_2
M_1	$m^2 + m\delta/p$	$m(1-m)$ $+m(1-\theta-m)\delta/p$
M_2	$m(1-m)$ $+(\theta-m)\delta/p$	$(1-m)^2 + (1-m)\delta/p$

Joint probabilities of transmission and nontransmission of M_1 and M_2 conditioned on QTL genotype Aa

Trans. vs nontrans.	M_1	M_2
M_1	$m^2 + \frac{m\delta(1-2p)}{2pq}$	$m(1-m) + \frac{m(1-\theta-m)\delta(1-2p)}{2pq}$
M_2	$m(1-m) + \frac{(\theta-m)\delta(1-2p)}{2pq}$	$(1-m)^2 + \frac{(1-m)\delta(1-2p)}{2pq}$

Joint probabilities of transmission and nontransmission of alleles M_1 and M_2 conditioned on QTL genotype aa

Trans. vs nontrans.	M_1	M_2
M_1	$m^2 - m\delta/q$	$m(1-m) - m(1-\theta-m)\delta/q$
M_2	$m(1-m) - (\theta-m)\delta/q$	$(1-m)^2 - (1-m)\delta/q$

Simulation details for univariate models

In the first step, the haplotypes (based on the QTL and the marker locus) are generated for each parent from a multinomial distribution with cell probabilities $(mp + \delta, mq - \delta, (1-m)p - \delta, (1-m)q + \delta)$. We note that δ can take values only between $\max\{-(1-p)(1-m), -(pm)\}$ and $\min\{p(1-m), m(1-p)\}$. If the generated haplotypes result in both parents being homozygous at the marker locus, we discard the data and repeat the process. In the second step, we generate the haplotype transmitted by each parent to the offspring based on the recombination fraction θ between the two loci. In the third step, we generate the QT value of the offspring conditioned on the QTL genotype resulting from the transmitted haplotypes. The trait value is generated from a specific distribution with mean α, γ or $-\alpha$ and variance σ^2 accordingly as the generated QTL genotype is AA, Aa or aa. Genotype data are generated for 500 pairs of parents, and if at least one parent is heterozygous within a pair, genotype as well as phenotype data are generated for the offspring. The simulation parameter values used are: $\alpha = 1$, $\gamma = 0, 1, 2$, $p = 0.3, 0.05$, $m = 0.5, 0.3, 0.1$ and $\theta = 0.01$. The empirical powers are computed based on 100,000 replications and level 0.05.

Simulation details for bivariate models

Suppose Y_1 and Y_2 are two QT controlled by a single biallelic QTL comprising alleles A and a with frequencies p and q, respectively. Assume that the joint distribution of (Y_1, Y_2) is bivariate normal.

Y_i can be expressed as $\mu_i + e_i, i = 1, 2$, where μ_i is the conditional mean of Y_i given the QTL genotype and is equal to $\alpha_i, 0$ or $-\alpha_i$, respectively, accordingly as the genotype is AA, Aa or aa, while e_i is the random error component with mean 0, variance ${\sigma_i}^2$ and the correlation between e_1 and e_2 is ρ, irrespective of the QTL genotype.

Then, the correlation between Y_1 and Y_2 can be obtained as:

$$\frac{\rho\sigma_1\sigma_2 + 2pq\alpha_1\alpha_2}{\sqrt{({\sigma_1}^2 + 2pq{\alpha_1}^2)}\sqrt{({\sigma_2}^2 + 2pq{\alpha_2}^2)}}.$$

We assumed $\alpha_1 = \alpha_2 = 1$ and $\sigma_1 = \sigma_2 = \sigma$ in our simulations. Thus, the above expression simplifies to $\frac{\rho\sigma^2 + 2pq}{\sigma^2 + 2pq}$. We evaluated the empirical powers of our method for four different values of the correlation coefficient between Y_1 and Y_2: 0, 0.2, 0.3 and 0.5. We used the simulation parameter values of $p = 0.3, 0.05$ and adjusted the values of σ and ρ to obtain the desired levels of correlation.

To simulate a bivariate phenotype comprising a QT and a binary trait, we first generate an observation from a bivariate normal distribution with the mean vector and the variance–covariance matrix conditioned on the QTL genotype as discussed above. We then dichotomize the second component of the generated bivariate observation by considering different thresholds. When $p = 0.3$, we used the threshold 2.2 resulting in a penetrance vector (0.266, 0.126, 0.048). When $p = 0.05$, we used the threshold 1.5 resulting in a penetrance vector (0.355, 0.132, 0.031).

Acknowledgements

This work was partially supported by the Council of Scientific and Industrial Research (CSIR) fellowship 09/093 (0111)/2008-EMR-I to Tanushree Haldar. The authors are grateful to Prof. Laura J. Bierut for continued collaboration with Saurabh Ghosh in the COGA project, through which the data on externalizing symptoms were available.

References

Abecasis G. R., Cardon L. R. and Cookson W. O. 2000 A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292.

Abecasis G. R., Cookson W. O. and Cardon L. R. 2001 The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am. J. Hum. Genet.* **68**, 1463–1474.

Allison D. B. 1997 Transmission-disequilibrium tests for quantitative traits. *Am. J. Hum. Genet.* **60**, 676–690.

Chen H. Y. and Li M. 2011 Improving power and robustness for detecting genetic association with extreme-value sampling design. *Genet. Epidemiol.* **35**, 823–830.

Ewens W. J., Li M. and Spielman R. S. 2008 A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker. *PLoS Genet.* **4**, e1000180.

George V., Tiwari H. K., Shu Y., Zhu X. and Elston R. C. 1999 Linkage and association analysis of alcoholism using a regression-based transmission/disequilibrium test. *Genet. Epidemiol.* **17 Suppl 1**, S157–S161.

Ghosh S., Bierut L. J., Porjesz B., Edenberg H. J., Dick D., Goate A. *et al.* 2008 A novel non-parametric regression reveals linkage on chromosome 4 for the number of externalizing symptoms in sib-pairs. *Am. J. Med. Genet.* **147**, 1301–1305.

Lange C. and Laird N. M. 2002 On a general class of conditional tests for family-based association studies in genetics: the asymptotic distribution, the conditional power, and optimality considerations. *Genet. Epidemiol.* **23**, 165–180.

Monks S. A. and Kaplan N. L. 2000 Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. *Am. J. Hum. Genet.* **66**, 576–592.

Rabinowetz D. 1997 A transmission disequilibrium tests for quantitative traits in nuclear families. *Hum. Hered.* **47**, 342–350.

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

Received 10 September 2014, in revised form 9 April 2015; accepted 15 May 2015
Unedited version published online: 26 May 2015
Final version published online: 30 October 2015